

GRAFF-MM : Graph-Frequency Filtering for Multi-Modal Memorability Prediction

Tahsin Tariq Banna

East West University

Swakshar Deb

University of Virginia

Shafin Rahman

North South University

Sejuti Rahman

`sejuti.rahman@newuu.uz`

New Uzbekistan University

Research Article

Keywords: Graph Convolutional Neural Network, GCN, Memorability, Multi-modal, Frequency Filter, Saliency

Posted Date: October 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10283943/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

GRAFF-MM : Graph-Frequency Filtering for Multi-Modal Memorability Prediction

Tahsin Tariq Banna^{1,2}, Swakshar Deb^{2,3}, Shafin Rahman⁴,
Sejuti Rahman^{2,5*}

¹Department of Computer Science & Engineering, East West University, Dhaka, Bangladesh.

²Department of Robotics & Mechatronics Engineering, University of Dhaka, Dhaka, Bangladesh.

³Department of Electrical & Computer Engineering, University of Virginia, Virginia, USA.

⁴Department of Electrical & Computer Engineering, North South University, Dhaka, Bangladesh.

^{5*}Department of Computer Science, New Uzbekistan University, Tashkent, Uzbekistan.

*Corresponding author(s). E-mail(s): s.rahman@newuu.uz;

Contributing authors: tahsin.banna@ewubd.edu;

swakshar.sd@gmail.com; shafin.rahman@northsouth.edu;

Abstract

Introduction: This study investigates the impact of modular components influencing image memorability, defined as a person’s ability to recall previously seen images or objects. Although memorability is influenced by various elements, the extent to which these factors contribute to overall memorability remains underexplored.

Methods: Existing models primarily focus on low-frequency information in visual scenes, often overlooking high-frequency data that may also capture traits contributing to image memorability. To address this, we introduce a novel Multi-Frequency Filter Bank Module (MFFBM) tailored for Graph Convolutional Networks (GCNs), enabling the extraction and utilization of low- and high-frequency information from graph representations of images and objects. Additionally, we explore multi-modal data sources such as visual saliency and image captions. Building on these components, we propose the Graph Frequency

Filtering Multi-Modal Memorability (GRAFF-MM) model, which integrates MFFBM with multi-modal data for enhanced memorability prediction.

Results: We evaluate our model on four benchmark datasets: LaMeM, SUN Memorability, OBJ Mem, and FIGRIM. The proposed model achieves competitive Spearman’s rank correlations of 0.692 on LaMeM, 0.711 on SUN Memorability, 0.653 on OBJ Mem, and 0.676 on FIGRIM, demonstrating its effectiveness in predicting image memorability.

Conclusion: We conduct an extensive ablation study to validate the effectiveness of the modules. We also visualize the learned features and edge cases to explain how our modules work in tandem to predict memorability.

Keywords: Graph Convolutional Neural Network, GCN, Memorability, Multi-modal, Frequency Filter, Saliency

1 Introduction

Images possess an intrinsic property called memorability, which quantifies the likelihood of an image being remembered or forgotten after it is viewed. This property reflects how well an image persists in a person’s memory and remains stable across individuals and observation delays [1]. Memorability has broad applications in fields such as education, advertising, design, and psychology, where creating impactful visuals is essential for effective communication and understanding human visual cognition. However, predicting image memorability is challenging, as humans are notably poor at accurately judging memorability scores [2]. This difficulty arises because memorability depends on various intrinsic and extrinsic factors, and human intuition about what makes an image memorable is often flawed and unreliable.

Several factors influence image memorability, including scene and environmental context [3, 4], recognizable faces [1, 5, 6], emotional attributes [5], the presence of distinct objects [7, 8], salient regions [9, 10], and attention [1, 11, 12]. Since human intuition about memorability is unreliable, researchers have turned to computational methods that predict memorability by extracting features from these factors. Early approaches using low-level, handcrafted features were limited in performance. However, deep learning methods have shown significant success in this domain [13–15].

Recent works have explored individual factors, such as attention [12], image captions [4], and spatial object information [16], to predict memorability. Although these approaches each contribute valuable insights, a comprehensive multi-modal strategy combining multiple factors has not yet been attempted. For example, attention models can highlight important regions in an image, but often fail to capture how objects within those regions interact. In contrast, spatial models focus on object relationships but may overlook the broader context. For example, in Fig 1, the spatial relation might infer that the relation between the motorcycle, the people, and the road is that the motorcycle is driving on top of it and the people are next to it. However, it may completely fail to understand that the motorcycle is going really fast and that people are standing by the fence to watch it go fast.

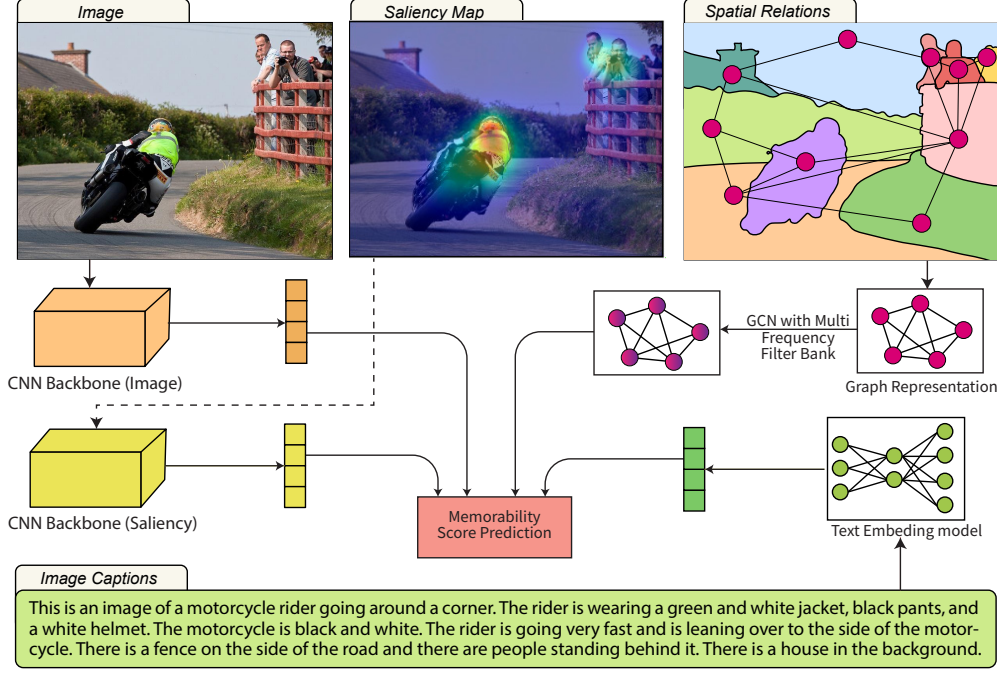


Fig. 1 Overview of the proposed multi-modal memorability prediction model. While existing models typically combine CNN backbones with image captioning, attention maps, or graph representations individually, ours is the first multimodal model to integrate all these components, while also introducing a novel module, the MFFBM.

Saliency significantly affects how visual information is encoded and retained in memory [17, 18]. Salient stimuli activate the hippocampus, a critical brain region for memory encoding and retrieval [19–21]. Text captions provide semantic context that aligns with cognitive schemas or trigger associations. Pairing visual stimuli with meaningful linguistic elements reinforces encoding, and narratives or descriptors associated with images often make them more engaging by invoking emotional or conceptual resonance [22–24]. So, in this work, we examine the different modalities that affect human memory and through which additional relations about the environment can be observed. In our investigations, we propose a multimodal architecture that utilizes spatial structures for object-centric feature encoding, image captioning for textual semantics, and saliency maps for contextual attention attributes, all of which combine, creating a state-of-the-art memorability prediction model. In order to process spatial structures, we extend the method proposed in [16] and use graph convolutional networks (GCN) along with graph filtering through a multifrequency filter bank. For image captioning features, we use an encoder from a Large Language Model (LLM) to find its feature embeddings. Finally, we extract saliency maps from the pre-trained DeepGaze IIE [25] model for weighted attention mapping.

The key contributions of our work are:

- We extend spatial structural architectures using graph spectral filtering through a multi-frequency filter bank that captures both low- and high-frequency information.
- We use LLM-based image captioning features for additional semantic understanding of the image and encode weighted contextual attention using extracted saliency maps to focus on regions with higher contributions.
- We propose a multimodal architecture that utilizes spatial structures, textual semantics, and contextual attention for better memorability prediction and perform in-depth ablation studies.

A preliminary version of this paper was published in GEMM [16], which only uses spatial structural information from objects. In the current version, we have broadly extended our previous work on several aspects: (a) we incorporate additional modalities to capture more comprehensive representations; (b) we incorporate LLM based captioning for deeper semantic understanding; (c) we introduce a novel frequency filtering module that captures a wider spectrum of information.

2 Related Work

Our methodology integrates novel components for feature extraction and model training. We detail the related work across these key areas, providing the necessary context for each of our proposed modules: attention, caption, saliency and spatial structure.

2.1 Quantification of Memorability

Isola et al. [1, 5] in 2011, introduced a method to quantify memorability scores of images through human participation, which involved participants looking at a set of *Target* and *Filler* images. The images were set to repeat after some delay, and the number of times a *Target* image was recognized on repetition was used to assign a score to the image. The more recognition the image had, the higher it scored. This method was conceptualized as a game called Visual Memorability Game and was distributed on the crowd-sourcing website, Amazon Mechanical Turk, where participants could play the game and provide the data to calculate the memorability scores. To ensure that the participants were focused on each of the images, the frequency of *Filler* image recognition was used. If this measure of attentiveness fell below some threshold, the data from the participants were discarded. This method of discarding inattentive data was named the "vigilance task". From this study, the first dataset on image memorability, the SUN Memorability Dataset, was created. Similarly, the LaMem [26], FIGRIM [3] and Object-Mem [7] data sets were created. We use these datasets in our experiments.

The core premise of the game involves presenting images to participants and observing their responses to repetitions of specific image types. These responses constitute the basis for assigning memorability scores to the images. Empirical studies on memorability show that it exhibits stability as a property, remaining constant over time [3] and invariant to contextual variations [27]. Furthermore, categorizing images by scenes and objects reveals that the memorability of images is predominantly influenced by the presence of objects within an image and their spatial relationships to each other.

Isola et al. [5] employed handcrafted features such as pixel histogram, GIST [28], SIFT [29], SSIM [30], and HOG [31–33] for predicting memorability scores.

2.2 Feature Engineering

Image memorability is intricately influenced by various local attributes and features, including spatial structures and local attributes within scenes. In 2012, Khosla et al. [8, 34] proposed a model that addressed the prediction of memorability specifically for image regions and introduced an image encoding process using a probabilistic framework. The rationale behind their model lies in the observation that individuals tend to retain certain objects and image regions more effectively than others.

Furthermore, Kim and Yoon et al. [9] presented a 2013 study introducing two novel spatial features that contribute to image memorability, taking into account both the size and position of objects within an image and the relative uncommonness of an object. Additional investigations [6, 26, 27] have indicated that scenes characterized by natural beauty or visual aesthetics tend to lack unusualness, thus exerting a negative influence on memorability. In addition, factors such as visual saliency [26, 34–36], image popularity [26], and emotions [14, 26] exhibit a positive correlation with image memorability. Our use of visual saliency as guidance for the model is directly influenced by the works of [26] & [9].

2.3 Learning Deep Features

Advancements in deep learning have significantly improved the ability of models to predict memorability compared to conventional hand-made models. A notable exemplar is MemNet, introduced by Khosla et al. in 2015 [26]. MemNet, a Hybrid-CNN model, was trained on the ILSVRC 2012 [37] and Places datasets. The authors also devised a baseline model that incorporates the features of HOG2x2, demonstrating superior performance, particularly in mitigating False Alarms (FA) and reducing signal noise. However, the study reported reduced performance when MemNet was fine-tuned on smaller datasets like SUN Memorability, emphasizing the need for a larger dataset, which led to the development of the Large-Scale Memorability Dataset (LaMem). Mancas et al. [11] studied the relations between saliency and memorability, finding that memorable images have unique localized regions and non-memorable images either lack unique regions or have multiple spread-out regions. In 2018, Fajtl et al. proposed AMNet [12], using transfer learning from deep models to improve predictions. Their approach used ResNet-50 [38] trained on ImageNet as the foundational classification model. The study revealed a correlation between strong performance in classification tasks and effective prediction of memorability, emphasizing the superiority of deep end-to-end learning features over specially designed visual features and those derived from CNN models. Later, Squalli-Houssaini et al. [4] introduced a novel deep learning model that combined CNN and Image Captioning (IC) features for image memorability prediction. Using a pre-trained VGG16 model on ImageNet, the study employed an encoder comprising CNN and LSTM components to create a 2D text-image embedding, enforcing alignment between the image and its corresponding semantic caption.

The method effectively predicted memorability through an aggregation step. In contrast, Perera et al. [39] proposed focusing solely on training a regression network on the final layer of a robust CNN model, using VGG16 as the base model. Their work demonstrated significant improvements in prediction scores. In this work we use deep networks for images and shallower networks for the other modules.

2.4 Human Vision Inspired Features

Visual saliency highlights the importance of certain regions or objects within a visual scene [40], grounded in the study of visual attention, which explores how the human visual system prioritizes environmental information. It identifies and quantifies the attention-grabbing attributes of elements in a visual stimulus. Given the overwhelming influx of visual information, the human brain employs selective attention to prioritize relevant details for cognitive processing. Visual saliency determines which regions or objects are most likely to attract attention, influenced by low-level features such as color, contrast, and orientation [41], and higher-level attributes like object semantics, context, and task relevance [3, 5, 7]. Low-level features drive bottom-up attention, where sensory distinctiveness automatically attracts focus. Highly saturated colors, sharp contrasts, and distinct edges are key contributors. Conversely, higher-level features influence top-down attention, with object semantics and context guiding focus toward meaningful or task-relevant elements, such as faces in social scenes. Computational models integrate these bottom-up and top-down mechanisms [20, 40] to predict attention allocation. By combining spatial and feature-based computations, these models generate saliency maps that quantitatively represent regions of interest, aiding in the analysis of perceptual significance across a scene.

2.5 Object Centric Spatial Features

There have been quite a few works on predicting image aesthetics using spatial structural information using graph neural networks [42, 43]. She et al. [42] use a hierarchical layout-aware GCN that can construct a graph in coordinate space and perform reasoning over spatial nodes. This is then aggregated into a latent space from which an aesthetic score can be predicted. Ghosal et al. [43] create a feature graph representation using a CNN backbone, which is used to train graph attention networks (GAT) to predict image aesthetics. Tahsin et al. [16] proposed the first method, GEMM, that predicted memorability using spatial structures. They constructed a fully connected graph from the object features in a region proposal network and trained GATs to predict image memorability. Our method includes a modified version of GCNs that ensures robustness.

3 Proposed Methodology

Previous methods isolate and analyze specific parts of the human visual process, that is, only semantics or saliency. This can lead the model to have a much more fragmented understanding of how memorable the images actually are with respect to their context. These methods mainly rely on CNN, GCN, and LSTM architectures [4, 16, 26] to

3.1 Problem Formulation

Consider a dataset $\mathbf{D}$ comprising images denoted by $\mathbf{I}_i$ with dimensions $W_i \times H_i \times 3$. The corresponding ground-truth memorability scores for each image are represented as $y_i \in [0, 1]$, which indicates the degree of memorability, where a higher score implies greater memorability. Let $\{(\mathbf{I}_i, y_i) : i \in [0, \mathcal{T}]\}$ be a set of tuples within the training dataset $\mathbf{D}$, where $\mathcal{T}$ is the total number of training images. The objective is to predict the memorability of an image accompanying its corresponding saliency map:

$$\hat{y}_j = \mathcal{M}(I_j, S_j, E_j; \theta), \quad \text{s.t.} \quad \hat{y}_j^{(k)} \approx y_j^* \quad (1)$$

Here, S_j and E_j denote the saliency map and text embedding associated with the image I_j , and θ are the model parameters.

The overall model $\mathcal{M}$ takes three different modalities as input, i.e., image, saliency, caption, and each modality is handled with three different modules, *(i)* Graph Convolutional module (Sec. 3.2.1), *(ii)* Visual Saliency (Sec. 3.2.2), *(iii)* Image Caption (Sec. 3.2.3). Different segments of an image contribute distinctively to its overall memorability, as indicated by empirical studies [8–10, 34]. Hence, similar to GEMM [16], we generate a dense graph from the image, $\mathcal{G} = (\mathbf{A}, \mathbf{V}, \mathbf{E})$ where $\mathbf{V}$ is a set of vertex, $\mathbf{E}$ is a set of edges and $\mathbf{A} \in \mathbb{R}^{N \times N}$ is the adjacency matrix of the graph $\mathcal{G}$ with N number of nodes. In our case, the vertices represent objects in an image, and we utilize the pre-trained Mask RCNN [44] to segment these regions. This graph-centric representation of an image $\mathbf{I}_i$ is given as input to our graph-convolutional module (Sec. 3.2.1). Apart from the image $\mathbf{I}_i$, to extract different modality-based information, we processed the saliency map and the image caption information through individual modules, Saliency Module and Image Caption module, respectively, which is elaborately discussed in Sec. 3.2.2 and 3.2.3. Finally, we concatenate their learned representations as the overall feature representation, which is then given as input to a regression head.

3.2 GRAFF-MM Architecture

The GRAFF-MM architecture, shown in Fig 2, uses multi-stage feature extraction processes to predict image memorability effectively. Images are preprocessed by resizing to a size of 250×250 px and randomly cropping and rotating about their centers before undergoing four distinct feature extraction stages: spatial, image, saliency, and captioning features. Here we discuss in detail the different modules for these features:

3.2.1 Multiple Frequency Filter Bank

Traditional graph convolution-based memorability prediction methods, such as the GEMM framework [16], have shown promise in extracting features from complex data structures such as images to predict memorability. The GEMM architecture primarily employs a Graph Attention Network (GAT) [45] to focus on object-centric viewpoints, aggregating information from neighboring nodes. However, this approach has a significant limitation: It relies predominantly on low-frequency information, potentially overlooking nuanced high-frequency signals that can carry critical context for image memorability (see Proposition 1). This focus on low-frequency content may lead to suboptimal representation, as high-frequency signals can be indicative of unique or

anomalous characteristics that can influence memorability. To address this limitation, we included a novel graph convolutional module namely “Multi-Frequency Filter-Bank Module” (**MFFBM**) that can be easily integrated with existing graph convolutional methods [16], without requiring any additional learnable parameters. Moreover, this MFFBM module does not include any extra computational burden such as eigendecomposition and higher order power of adjacency matrix. Furthermore, our MFFBM module significantly enhances traditional graph-based frameworks [16] by leveraging both low- and high-frequency information to create a more comprehensive representation for image memorability prediction. We first construct a low-pass filter which is similar to traditional GCN filters such as GEMM [16], GCN [46].

The low-pass filter primarily aggregates a node’s self-information with its neighborhood information, representing a low-pass filtering operation. Mathematically, this low-pass filtering is described by the following equation:

$$\begin{aligned}\mathbf{H}_{\text{low}}^{(\ell+1)} &= \left(\tilde{\mathbf{A}} \odot \mathbf{M}^{(\ell)} \right) \mathbf{H}^{(\ell)} \mathbf{W}^{(\ell)} \\ &= \mathcal{A}_m \mathbf{H}^{(\ell)} \mathbf{W}^{(\ell)}\end{aligned}\quad (2)$$

where $\tilde{\mathbf{A}}$ is the graph adjacency matrix with self-loops, $\mathbf{M}^{(\ell)}$ is an attention mask at the ℓ -th layer, $\mathcal{A}_m$ is the masked adjacency, and $\mathbf{H}^{(\ell)}$ represents the node features at layer ℓ .

This low-pass filter captures the general structure (e.g. similarity between objects features) of the graph, which provides a meaningful representation for memorability prediction [16]. Eqn. 2 is also commonly used by the authors of **GEMM** [16] for extracting low-frequency features.

To complement the low-pass filter, we introduce a high-frequency-emphasizing filter designed to capture high-frequency signals that represent dissimilarity among neighboring nodes. This filter computes the difference between a node’s self-information and its scaled aggregated neighborhood information, emphasizing distinctive node characteristics. It is mathematically expressed as follows:

$$\begin{aligned}\mathbf{H}_{\text{high}}^{(\ell+1)} &= \mathbf{H}^{(\ell)} \mathbf{W}^{(\ell)} - \alpha \mathcal{A}_m \mathbf{H}^{(\ell)} \mathbf{W}^{(\ell)} \\ &= (\mathbf{I} - \alpha \mathcal{A}_m) \mathbf{H}^{(\ell)} \mathbf{W}^{(\ell)}\end{aligned}\quad (3)$$

where $\alpha \in [0, 1]$ controls how much neighborhood information is deducted from the self-features. Higher values of α emphasize dissimilarity between a node and its neighbors, producing stronger high-frequency-emphasizing behavior, while lower values retain more of the original self-features.

Although the low- and high-frequency-emphasizing components acquired through Eqns. 2 & 3 capture complementary aspects of the graph signal, neither alone provides a complete representation. Hence, we aggregate this information together to obtain the overall representation as follows:

$$\mathbf{H}^{(\ell+1)} = \beta \mathbf{H}_{\text{high}}^{(\ell+1)} + (1 - \beta) \mathbf{H}_{\text{low}}^{(\ell+1)} \quad \in \mathbb{R}^{N \times d_{GCN}} \quad (4)$$

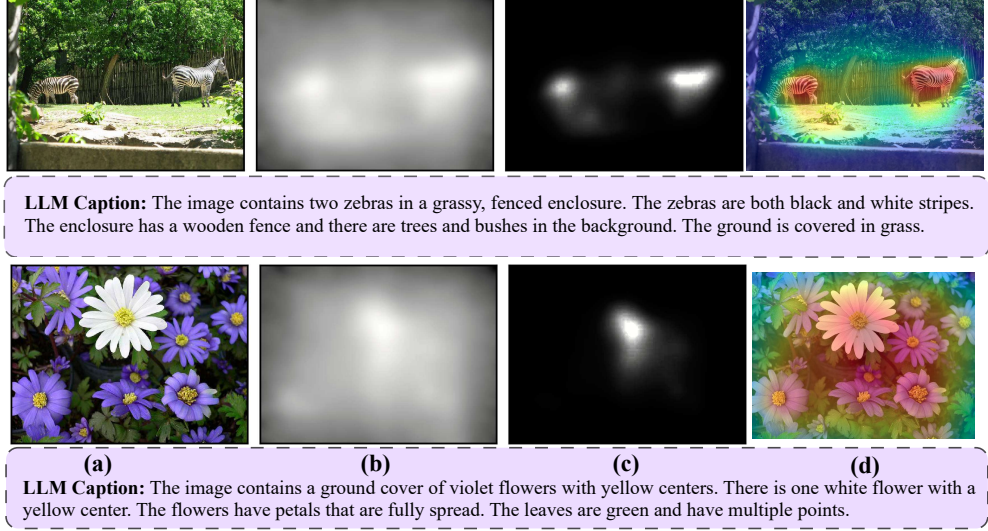


Fig. 3 The image saliency map along with the caption generated by **DeepGaze** [25] and **Gemini LLM** respectively, that is utilized in this work. (a) Input images from the **LaMem** dataset. (b) Eye fixation maps generated by **DeepGaze** [25]. (c) Applying soft threshold to the original fixation maps. (d) Heatmap generated from the visual saliency (b).

where the hyperparameter β serves as a balancing factor, controlling the weight between the low- and high-frequency contributions (See Proposition 2). Finally, to aggregate information across all layers, we define the final graph representation as a skip-sum over intermediate layer outputs:

$$\mathbf{H}_{\text{out}} = \sum_{i=0}^{L-1} \mathbf{H}^{(i+1)} \quad (5)$$

This skip-sum aggregation helps retain information from intermediate graph representations. Together with the two-branch MFFBM formulation in Eqns. 3–4, the model can incorporate frequency-emphasizing graph features beyond the single-branch low-pass graph convolution used in GEMM [16]. Furthermore, Eqns. 3–4 can be easily integrated into existing graph-based frameworks while improving the overall representation for image memorability prediction.

3.2.2 Visual Saliency Module

Beyond solely focusing on object-centric features, we also include an image saliency map [25], which strongly correlates with human gaze patterns. Prior work has shown that saliency maps generated from visual-attention models can provide useful information when compared with eye-fixation maps for eye-tracking analysis [47]. This saliency map can uncover significant intrinsic cues in a given image [34] and offers a

more comprehensive perspective to predict memorability through a holistic representation of features. In general, people naturally gaze at their focus with different levels of intensity on each part of the image, which is largely due to our intrinsic attention behavior toward the most important and related parts. This intrinsic quirk of the human attention mechanism is expressed through visual saliency and is a very good predictor of object memorability in simple scenes [7]. To identify important regions of an image, we compute the visual saliency map using a pre-trained DeepGaze IIE model [25]. The saliency map is the same size as the image, with pixel values indicating the weighted intensity of gaze, ranging from 0 to 1 (normalized). We note that the resultant saliency map is very spread out, although saliency and, to some extent, memorability, exhibit a center bias. This center bias results from the photographer’s bias and viewing strategy [48–50]. Experimentally, we find that a centrally biased saliency map is better at predicting memorability scores.

If we denote the saliency map of an image as S then we calculate the weighted center of all saliency values in the image with coordinates (x, y) given as:

$$x = \frac{\sum_{i,j} i * S[i][j]}{\sum_{i,j} S[i][j]}; y = \frac{\sum_{i,j} j * S[i][j]}{\sum_{i,j} S[i][j]}$$

Here, i and j represent pixel coordinates within the image, and $S[i][j]$ represents the corresponding saliency value of that pixel. This calculation essentially finds the center of mass for the saliency distribution within the image.

To focus on the salient objects, we apply a transformation, τ , on the returned saliency map to get the output map $\acute{S}$. Instead of a hard threshold, $\tau(S)$, defined,

$$\acute{S} = \tau(S) = \begin{cases} 1, & \text{if } S(x, y) > T \\ 0, & \text{if } S(x, y) \leq T \end{cases} \quad (6)$$

We apply a soft threshold, $\tau_s(S)$, defined as,

$$\acute{S} = \tau_s(S) = cS(x, y)^\gamma \quad (7)$$

We opt for a soft threshold as it accommodates temporal and individual variations, even when the saliency of an object remains constant. This enhances the robustness of our model to such variations, while still prioritizing the most salient region in the image, as illustrated in Fig 3.

After applying a soft threshold, we get the saliency map $\acute{S}$. We multiply this map with the original to get selective saliency-based attention and learn an intermediate representation through a set of convolutional operations that we define as the feature extractor $f_s(\cdot)$ to get the representation $\mathbf{R}_s$,

$$\mathbf{R}_s = f_s(\acute{S}); \quad \in \mathbb{R}^{B \times d_s} \quad (8)$$

where, d_s is the output dimension of the feature extractor and B is the batch size.

3.2.3 Image Captioning Module

Although the **MFFBM** framework captures the relationships between objects in a scene that can recognize contextual mismatch (i.e., unusualness), it can ignore local contextual information in a scene, such as emotions (e.g., anger, happiness) or actions, which can further influence the memorability score. Therefore, to incorporate this information from the scene, we explored using image captions as additional semantic information for the images (see Figure 3). We use a multimodal generative large language model for extracting the image captions and a text embedding model to get the embeddings from those captions. Specifically, we use the Google Gemini LLM and the gecko text embedding model from the vertex LLM SDK. This gives us the embedding text E_c .

3.2.4 Prediction module

The representation for the image I is extracted using an **AMNet** [12] based feature extractor $f_i(\cdot)$. We use the function $f_c(\cdot)$ on the text embedding E_c to learn a mapping to the appropriate dimension in order for concatenation. We concatenate the image, saliency, and text caption representations and pass it through some fully connected layers to get the visual-semantic prediction, $\mathbf{Y}_{isc}$,

$$\mathbf{Y}_{isc} = F_{MLP}(f_i(I) \otimes f_s(\hat{S}) \otimes f_c(E_c)); \quad \mathbf{Y}_{isc} \in \mathbb{R}^{B \times 1} \quad (9)$$

On the other hand, we use the final graph representation, $\mathbf{H}_{out}$, to obtain the spatio-contextual prediction $\mathbf{Y}_g$ using the function $\vec{h}_i(\cdot)$:

$$\mathbf{Y}_g = \vec{h}_i(\mathbf{H}_{out}), \quad \mathbf{Y}_g \in \mathbb{R}^{B \times 1}. \quad (10)$$

Our final memorability prediction, $\hat{\mathbf{Y}}$ is a weighted sum of $\mathbf{Y}_{isc}$ and $\mathbf{Y}_g$,

$$\hat{\mathbf{Y}} = C \times \mathbf{Y}_{isc} + (1 - C) \times \mathbf{Y}_g, \quad \hat{\mathbf{Y}} \in \mathbb{R}^{B \times 1} \quad (11)$$

where C is the balancing weight. This is necessary to fuse $\mathbf{Y}_{isc}$ and $\mathbf{Y}_g$ together and to ensure that neither of them dominates and skews the overall prediction.

The GCN with the **MFFBM** is trained using the MSE loss function $\mathcal{L}_{gcn} = (\hat{y} - y)^2$.

The contextual and semantic net is trained similarly to the loss implemented in [12], where the first term is the MSE loss function and the second term is a weighted penalty of $l_1 - l_2$ regularization, $\mathcal{L}_{csn} = \mathcal{L}_{gcn} - \gamma \mathcal{L}_\alpha$, where $\gamma \geq 0$ is a scalar hyperparameter controlling the contribution of the penalty term $\mathcal{L}_\alpha$.

3.2.5 Inference

Before inference, the image’s graph representation, visual saliency, and textual description are extracted. These data are then fed into the model, where the image and saliency features are processed through separate convolutional blocks. The caption is transformed into an embedding, and its features are extracted. Meanwhile, the

graph representations pass through the GCN with MFFBM block to obtain additional features. All of these features are then aggregated to make the final prediction.

Computation Complexity: MFFBM is highly computationally efficient. The computation time complexity is similar to GAT [45] as $\mathcal{O}(Nd^2 + |\mathbf{E}|d)$, here, $|\mathbf{E}|$ is the total number of edges. Due to the natural sparsity of graphs, the message propagation yields complexity of $\mathcal{O}(|\mathbf{E}|d)$ with PyTorch geometry [51].

3.3 Theoretical Analysis

Lemma 1 The frequency profile of any given graph convolution kernel $\mathbf{C}$ can be defined in the spectral domain by the filter response:

$$\mathcal{F}(\lambda) = \text{diag}^{-1} \left(\mathbf{U}^T \mathbf{C} \mathbf{U} \right)$$

where, $\mathcal{F}(\cdot)$ is the frequency response, $\mathbf{C}$ is the convolutional kernel and $\mathbf{U}$ the eigenvector matrix of the normalized graph Laplacian $\mathbf{L}$.

Proposition 1 The frequency profile of the graph convolution layer utilized in **GEMM**[16] is defined by:

$$\mathcal{F}(\lambda) = 1 - \frac{p}{p+1} \lambda \quad (12)$$

where λ is the eigenvalues of the normalized graph Laplacian and the given graph is an undirected regular graph whose node degree is p .

The graph constructed in the **GEMM** architecture is a dense graph. Therefore, the degree of the node, p , is the same for each of the nodes. This demonstrates that the **GEMM**[16] frequency profile acts as a low-pass filter. Since the given graph is an undirected dense graph, all node degrees are equal to a number of object regions N (i.e., extracted through Mask RCNN [44]), leading to a frequency profile defined by $1 - \frac{N\lambda}{N+1}$. Moreover, the normalized graph Laplacian eigenvalues are in the range $[0, 2]$ and the magnitude of the filter linearly decreases until the third quarter of the spectrum (cutoff frequency), where it reaches zero. Then it linearly increases until the end of the spectrum. Due to page limitation, we omit this proof, and readers are encouraged to follow Balcilar *et al.* [52] for the detailed proof of the proposition 1.

Proposition 2 Under a linearized setting with a fixed symmetric normalized graph operator, the proposed MFFBM in Eqn. 4 extends the baseline low-pass graph convolution to a parameterized family of constrained polynomial graph filters. The resulting spectral response is governed by α and β , enabling frequency-emphasizing behaviours beyond the fixed low-pass response of the baseline.

Proof For the spectral analysis, we consider a linearized setting in which nonlinear activation functions are omitted and $\mathcal{A}_m$ is treated as a fixed symmetric normalized graph operator

across layers. According to Eqn. 4, the MFFBM update at layer ℓ is:

$$\mathbf{H}^{(\ell+1)} = \beta(\mathbf{I} - \alpha\mathcal{A}_m)\mathbf{H}^{(\ell)}\mathbf{W}^{(\ell)} + (1 - \beta)\mathcal{A}_m\mathbf{H}^{(\ell)}\mathbf{W}^{(\ell)} \quad (13)$$

Since $\mathbf{W}^{(\ell)}$ acts on feature channels, the node-domain spectral response is determined by the polynomial terms in $\mathcal{A}_m$.

We can write $\mathbf{H}^{(1)}$ by combining $\mathbf{H}_{\text{low}}^{(1)}$ and $\mathbf{H}_{\text{high}}^{(1)}$:

$$\mathbf{H}^{(1)} = [\beta\mathbf{I} + (1 - \beta - \beta\alpha)\mathcal{A}_m]\mathbf{H}^{(0)}\mathbf{W}^{(0)} \quad (14)$$

To illustrate the induced second-order spectral response, consider the second-layer representation obtained by applying Eqn. 4 again. We can express the low-frequency part of the second layer as $\mathbf{H}_{\text{low}}^{(2)} = \mathcal{A}_m\mathbf{H}^{(1)}\mathbf{W}^{(1)}$, which can be further decomposed as:

$$\mathbf{H}_{\text{low}}^{(2)} = [\beta\mathcal{A}_m + (1 - \beta - \beta\alpha)\mathcal{A}_m^2]\mathbf{H}^{(0)}\mathbf{W}^{(0)}\mathbf{W}^{(1)} \quad (15)$$

where $\mathbf{C}_1 = \beta\mathcal{A}_m + (1 - \beta - \beta\alpha)\mathcal{A}_m^2$ is the corresponding graph filtering operator. We can also express this operator with the help of the graph filter response function $\mathcal{F}(\cdot)$ as $\mathbf{C}_1 = U\text{diag}(\mathcal{F}_1(\lambda))U^T$ [52]. Hence, we analyse the filter response $F_1(\lambda)$ of $\mathbf{H}_{\text{low}}^{(2)}$ as $F_1(\lambda) = \text{diag}^{-1}(U^T\mathbf{C}_1U)$. Using the relation $\mathcal{A}_m = \mathbf{I} - \mathbf{L}$ under the fixed symmetric normalized graph-operator assumption, the filter response $F_1(\lambda)$ for $\mathbf{H}_{\text{low}}^{(2)}$ is:

$$F_1(\lambda) = \beta(1 - \lambda) + (1 - \beta - \beta\alpha)(1 - \lambda)^2 \quad (16)$$

Now we expand the high-frequency-emphasizing branch of Eqn. 4:

$$\begin{aligned} \mathbf{H}_{\text{high}}^{(2)} &= [\beta\mathbf{I} + (1 - \beta - 2\beta\alpha)\mathcal{A}_m \\ &\quad - \alpha(1 - \beta - \beta\alpha)\mathcal{A}_m^2]\mathbf{H}^{(0)}\mathbf{W}^{(0)}\mathbf{W}^{(1)} \end{aligned} \quad (17)$$

where $\mathbf{C}_2 = \beta\mathbf{I} + (1 - \beta - 2\beta\alpha)\mathcal{A}_m - \alpha(1 - \beta - \beta\alpha)\mathcal{A}_m^2$ is the corresponding graph filtering operator for $\mathbf{H}_{\text{high}}^{(2)}$. Using the relation $\mathcal{A}_m = \mathbf{I} - \mathbf{L}$, the filter response $F_2(\lambda)$ is:

$$F_2(\lambda) = \beta + (1 - \beta - 2\beta\alpha)(1 - \lambda) - \alpha(1 - \beta - \beta\alpha)(1 - \lambda)^2 \quad (18)$$

Finally, combining the two branch responses according to Eqn. 4 gives the overall frequency profile:

$$\begin{aligned} F(\lambda) &= \beta F_2(\lambda) + (1 - \beta)F_1(\lambda) \\ &= [\beta + (1 - \beta - \beta\alpha)(1 - \lambda)]^2 \end{aligned} \quad (19)$$

Thus, the two-layer linearized MFFBM induces a constrained quadratic polynomial spectral response. When $1 - \beta - \beta\alpha \neq 0$, the unconstrained minimum of this response occurs at

$$\lambda^* = 1 + \frac{\beta}{1 - \beta - \beta\alpha}. \quad (20)$$

If λ^* lies outside the valid normalized Laplacian spectrum $\lambda \in [0, 2]$, the minimum over the spectrum occurs at one of the boundary frequencies. This shows that α and β control the shape and location of the spectral response. More generally, repeated linearized applications of the MFFBM update yield higher-order constrained polynomial responses in $\mathcal{A}_m$. Therefore, the linearized MFFBM induces a controllable family of frequency-emphasizing polynomial responses beyond the fixed low-pass response of GEMM. $\square$

Table 1 Summary of key properties of datasets commonly used for image memorability prediction.

Feature	LaMem	SUN Mem	FIGRIM	Obj Mem
Source	Diverse datasets	SUN[31]	SUN[31]	PASCAL-S
Images	58,741	2,222	1,754	850
Train	45,000	1,111	877	680
Test	10,000	1,111	877	170
Validation	3,741	-	-	-
Resolution	Varied	256x256px	700x700px	Varied

4 Experiments

4.1 Setup

Datasets: We employed four datasets, which have been widely utilized by current state-of-the-art methods in the field. These datasets include the LaMem dataset [26], the SUN memorability dataset [1], the FIGRIM dataset [3], and the Object Memorability Dataset [7] (Summary of the datasets are shown in Table 1). The four data sets were generated using variations of the original memorability game, in which data collection involved human participants, and the scores were annotated based on the statistical performances of these participants.

Evaluation Metrics: Image memorability prediction is a supervised regression task that is typically evaluated using metrics such as mean absolute error (MAE), mean squared error (MSE) and root mean squared error (RMSE). Many studies, such as [5, 26], favor Spearman rank correlation (RC), which measures monotonic relationships and emphasizes the relative order of the data points. Unlike Pearson correlation, Spearman’s RC captures nonlinear monotonic trends. Furthermore, RMSE is often used, as in [1, 12], to assess model errors and assess overfitting or underfitting, helping to gauge generalization performance.

Implementation details: We used the ADAM [53] optimizer with a fixed learning rate of 0.001. Training was done using mini-batches of 64 images with early stopping enabled when the validation loss was at its minimum. Due to the large size of the LaMem dataset, we implemented a custom dataloader to store the extracted feature locations in memory, loading them dynamically during runtime. The model was trained for 150 epochs, with early stopping typically triggering between 75 and 90 epochs, depending on the dataset. All experiments were conducted on a single Nvidia RTX 3080 GPU using the PyTorch framework.

4.2 Quantitative Results

Table 2 provides a comparative performance analysis of our proposed model against existing approaches across various datasets. The evaluation metrics include Spearman’s Rank Correlation (RC) and Root Mean Squared Error (RMSE). A higher RC indicates better alignment with human judgment, while a lower RMSE reflects higher prediction accuracy.

Table 2 Performance of the proposed models compared to existing models. The highest model performances are highlighted in **bold** and second best in bold gray.

Method\Metrics	Spearman's Rank Correlation (RC) $\uparrow$				RMSE ($\times e^{-2}$) $\downarrow$			
	FIGRIM	OBJ mem	SUN	LaMeM	FIGRIM	OBJ mem	SUN	LaMeM
AMNet [12]	0.490	0.540	0.650	0.677	1.570	0.674	1.100	0.820
MemBoost [39]	0.570	-	0.660	0.670	-	-	-	-
ICNet [4]	0.480	-	0.590	0.720	-	-	-	-
Dubey et al [7]	-	0.390	-	-	-	-	-	-
MemNet [26]	0.425	0.513	0.630	0.640	0.735	1.851	-	-
GEMM [16]	0.580	0.597	0.690	0.706	1.350	0.791	2.130	0.750
Human Consistency	0.740	0.760	0.750	0.680	-	-	-	-
GRAFF-MM (Ours)	0.676	0.653	0.711	0.692	1.222	0.584	0.985	0.808

MemNet [26], one of the pioneering models for memorability prediction, demonstrates limited performance. This can be attributed to its reliance on basic CNNs, which do not capture complex high-level visual features.

Dubey et al. [7] introduced the OBJMem dataset with object segmentation and used an SVR to extract deep features. However, their approach performs poorly because of its limited capacity to effectively utilize object-level information for memorability prediction. Our approach, on the other hand, integrates a more robust feature extraction through RPN, enabling precise visual and contextual modeling. ICNet [4] achieved strong performance in the LaMem dataset, attributed to its use of image captions for memorability prediction. However, it struggles with datasets such as FIGRIM and SUN-Mem, which contain numerous similar images from the SUN dataset. This decline stems from the use of captions generated by [54], which tend to be generic and do not provide the necessary variability for closely related images. In contrast, the captions generated for our model are semantically richer and more descriptive. This allows our approach to better differentiate among similar images, resulting in superior performance. AMNet [12] employs neural attention mechanisms to estimate memorability. Although effective, its design does not align fully with the way humans perceive and process visual information. Our model incorporates visual attention, which better reflects human visual cognition, leading to better alignment with human judgments and improved accuracy in the prediction of memory. GEMM [16] leverages spatial relationships using GCNs to enhance memorability prediction. Many standard neighborhood-aggregation based GCNs mainly behave as smoothing operators, emphasizing low-frequency graph signals. This is useful for capturing broad object-level structure, but memorability can also depend on distinctive spatial and object-level cues, such as uncommon object arrangements, salient regions, and local deviations from surrounding context. The proposed MFFBM therefore combines low-frequency and high-frequency graph responses, enabling the model to represent both broad spatial structure and distinctive object-relation cues for memorability prediction. Our analysis shows that AMNet [12] has 38.8 million parameters, while GEMM [16] has 39.9 million. Our proposed GRAFF-MM model uses 40.1 million parameters, incurring only a marginal overhead.

Table 3 Spearman’s Rank Correlation across various GCN layers and loss functions, on the OBJ mem dataset. Here, RC_{MSE} and RC_{Huber} denote scores when trained with MSE and Huber loss, respectively.

GCN layers	GRAFF-MM		GEMM	
	$RC_{MSE} \uparrow$	$RC_{Huber} \uparrow$	$RC_{MSE} \uparrow$	$RC_{Huber} \uparrow$
1	0.583	0.598	0.581	0.574
2	0.653	0.628	0.597	0.578
3	0.623	0.601	0.575	0.583
4	0.600	0.608	0.577	0.585
5	0.614	0.608	0.577	0.582

4.3 Ablation Study

Effect of Layers: Table 3 presents an ablation study on the influence of GCN layers, where the number of layers was incrementally increased from 1 to 5, keeping the rest of the architecture unchanged. Performance was evaluated using Spearman’s rank correlation on test sets. Both our method and GEMM [16] showed improved performance from 1 to 2 layers due to the ability to capture more complex features. Beyond 2 layers, generalization declined as overfitting occurred, with training performance improving but validation performance plateauing or decreasing. Our method consistently outperformed GEMM starting from layer 2, thanks to its filterbank-based graph convolutional approach, which captures a broader frequency range and identifies subtle patterns contributing to image memorability. In contrast, GEMM relies on low-pass filtering, which restricts its ability to extract complex patterns as layers deepen. Although both methods performed similarly at layer 1, our approach maintained superior performance across layers, highlighting its adaptability to diverse features, while GEMM became more restrictive with deeper layers.

Analysis of α and β : In our graph filter-bank approach, $\alpha \in [0, 1]$ controls the scale of the aggregated neighborhood term subtracted from the self-features in the high-frequency-emphasizing branch. Larger values of α increase the difference between a node representation and its aggregated neighborhood representation, producing a stronger high-frequency response, while smaller values preserve more of the original self-features. The parameter β controls the balance between the high-frequency and low-pass branches in the output $\mathbf{H}^{(\ell+1)}$. Larger values of β assign more weight to the high-frequency branch, whereas smaller values prioritize the low-pass branch.

Fig. 4 shows the effect of β on the x-axis and α on the y-axis on rank-correlation performance. The best observed result occurs at $\beta = 0.5$ and $\alpha = 0.6$, achieving a Spearman’s rank correlation of 0.67. This suggests that, for the FIGRIM dataset, a moderate neighborhood subtraction strength combined with a balanced contribution from both branches is beneficial. The setting $\alpha = 0.6$ indicates that moderately scaling the aggregated neighborhood term helps emphasize node distinctiveness while still preserving structural context. Similarly, $\beta = 0.5$ suggests that combining the low-pass and high-frequency branches captures complementary graph information.

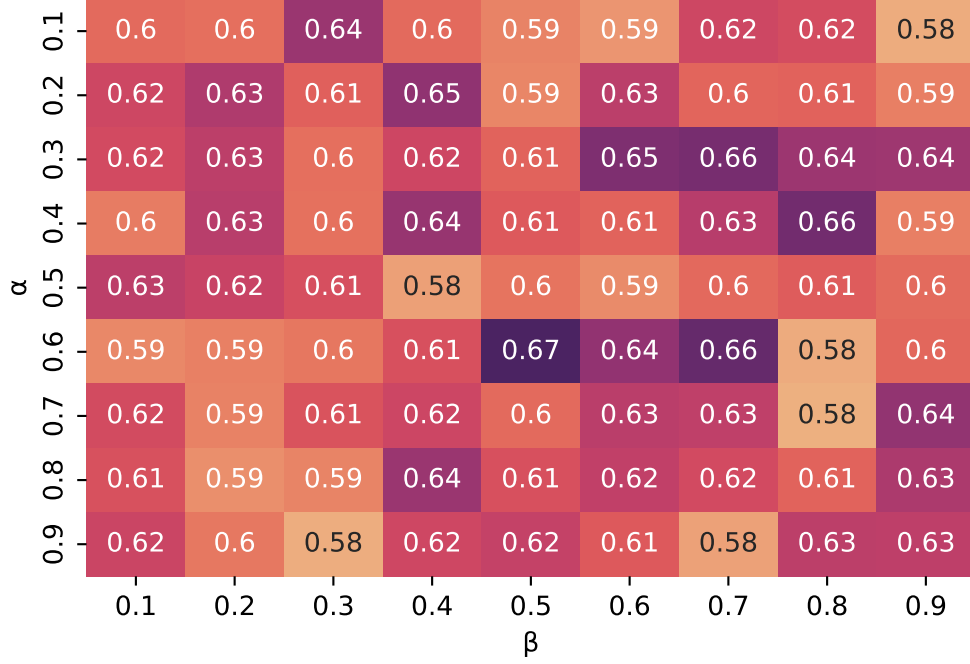


Fig. 4 Spearman’s Rank Correlation for the hyperparameters β and α , on the FIGRIM dataset.

Table 4 Incremental performance improvements over the baseline GEMM, achieved by incorporating the MFFBM, Saliency, and IC modules. It highlights the contribution of each module to the overall enhancement of the model’s performance.

Included Modules			Spearman’s Rank Correlation (RC) $\uparrow$				RMSE ($\times e^{-2}$) $\downarrow$			
MFFBM	Saliency	IC	FIGRIM	ObjMem	SUN	LaMeM	FIGRIM	ObjMem	SUN	LaMeM
-	-	-	0.50	0.42	0.61	0.62	1.363	0.656	1.370	0.813
✓	-	-	0.563	0.632	0.702	0.618	1.234	0.598	0.953	1.043
-	✓	-	0.613	0.575	0.708	0.709	1.158	1.353	0.970	0.770
-	-	✓	0.662	0.573	0.692	-	1.222	0.650	1.028	-
-	✓	✓	0.665	0.586	0.692	-	1.354	0.595	1.013	-
✓	✓	-	0.631	0.576	0.712	0.682	1.360	1.353	0.963	0.820
✓	-	✓	0.662	0.573	0.692	0.700	1.237	0.650	1.028	0.783
✓	✓	✓	0.676	0.653	0.711	0.692	1.222	0.584	0.985	0.808

The heatmap also shows relatively stable performance across several mid-range values of α and β , while more extreme combinations tend to produce weaker results. This indicates that the model benefits from balancing neighborhood subtraction and branch weighting, rather than relying too heavily on either the low-pass or high-frequency component alone.

Effect of modules: In our baseline, the model discriminates between images with different memorability scores using only the graph-based spatial representation. This is effective because the graph representation captures object-level spatial relations

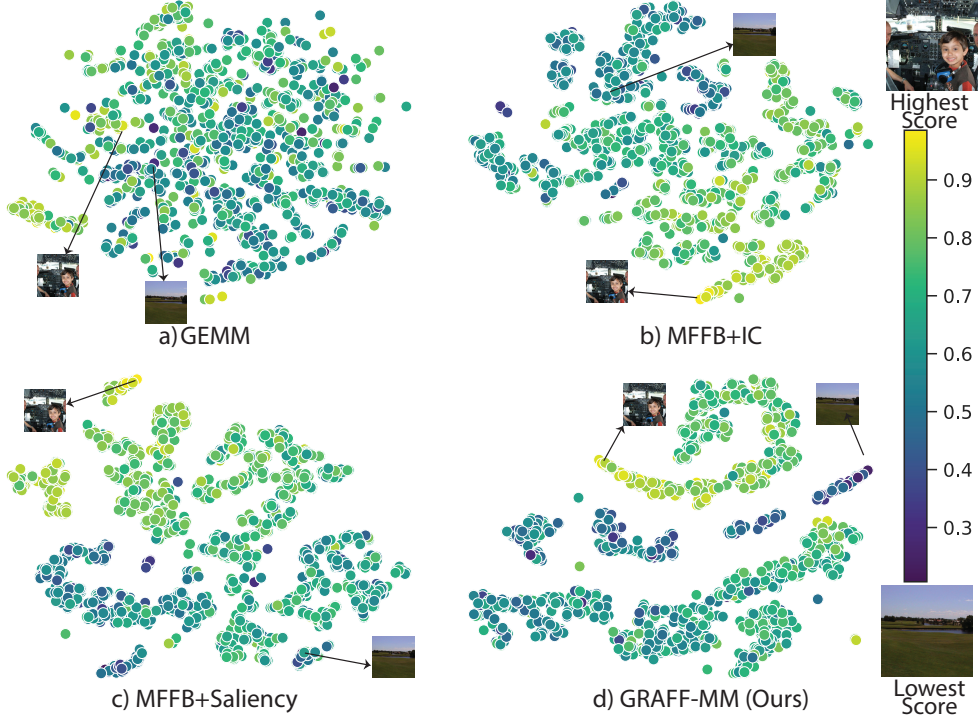


Fig. 5 t-SNE visualizations of features learned from the **FIGRIM** dataset: **(a)** Features learned by the baseline GEMM model. **(b)** Incorporation of the MFFBM along with IC. **(c)** Addition of the MFFBM with saliency. **(d)** Features learned by our proposed model, which integrates both IC and saliency with the MFFBM.

within the image. However, the model can be further improved by providing additional information that is not fully captured by either the raw image features or the graph representation alone. Hence, as shown in Table 4, introducing saliency and image-captioning features improves the baseline by adding attention-guided visual cues and semantic contextual information.

Furthermore, as discussed in Proposition 1 and 2, the baseline graph convolution is restricted to a fixed low-pass response, whereas MFFBM induces a controllable family of frequency-emphasizing polynomial responses. This allows the model to incorporate a broader range of graph spectral information beyond single-branch low-pass aggregation. The MFFBM module improves performance on FIGRIM, OBJMem, and SUN, indicating that frequency-emphasizing graph features provide useful complementary information for memorability prediction.

The full model, which combines MFFBM, saliency, and image-captioning modules, achieves the most balanced overall performance across the benchmark datasets. This suggests that the three modules capture complementary memorability cues: MFFBM captures graph spectral variations, saliency provides attention-guided visual information, and image captions add semantic context that may not be fully represented by object-level graph relations alone.

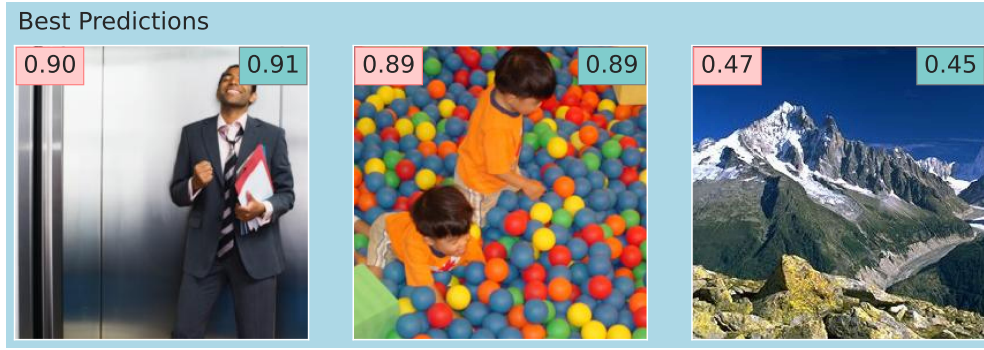
One possible trade-off of MFFBM is that its spectral response is controlled by scalar hyperparameters α and β . For visually diverse datasets such as LaMem, a single frequency-emphasis setting may interact differently with different graph structures, making the contribution of MFFBM more sample-dependent than in smaller or more homogeneous datasets.

Feature Visualization: To demonstrate the effectiveness of our proposed method, we performed t-SNE [55] visualizations of feature representations learned from the FIGRIM memorability dataset. Fig 5 presents a comparison of the baseline GEMM model [16], its enhanced variants (MFFBM + IC and MFFBM + saliency) and our proposed approach. The baseline GEMM model exhibits significant limitations, as reflected in the scattered clusters, where points with identical labels (represented by similar colors) are widely dispersed with significant overlap between points of different labels. The incorporation of the MFFBM module along with image captioning improves the cohesiveness of the clusters, leading to more distinct groupings, although some overlap remains. Similarly, integrating MFFBM with saliency further improves cluster separability. The addition of the different modules provide more concrete information for the model to process and in effect, improves feature representation.

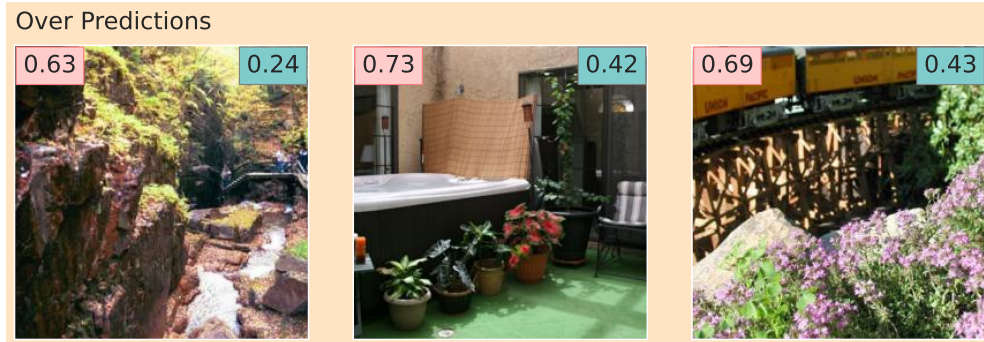
4.4 Visualization

Our proposed model, which combines MFFBM with both image captioning and saliency, achieves the best feature representation, as evident from the well-separated and compact clusters. The discernible clusters of similar data-points means that the model is able to discriminate well between images with close memorability scores as it is evident from the baseline t-SNE. This validates the effectiveness of our approach and shows the importance of combining multi-scale feature fusion, saliency information, and image captioning.

Extreme cases: In this section, we explore some instances of failure cases, i.e. over and under prediction, of our model when predicting the image memorability score (0-1, higher score indicates better memorability). Over-prediction occurs when the model’s predicted memorability score for an image is higher than the original label. In contrast, under-prediction occurs when the model’s predicted score is lower than the original label. In Fig 6, we observe that under-prediction typically occurs in our model for images that contain potentially memorable elements, such as humans, that are shaky or out of focus. This lack of clarity can lead to a lower predicted score as the model may struggle to identify these elements accurately or assess their contribution to overall memorability. However, over-predictions tend to occur in images with bland scenes or scenes with various semantic content. The richness in visual elements might cause the model to overestimate the memorability score, as it may sometimes incorrectly associate diversity within an image as a sign of memorability. Some under predicted images share a unique feature similar to low memorability images where there is a lack of significantly salient object. On the other hand, over-predicted images share features common to highly memorable images, such as local clusters of vibrant colors that may increase perceived object importance. Recognizing that our model can under-predict when encountering images with notable but unclear elements and over-predict when faced with complex scenes, it can be useful for future researchers in this domain.



(a) Closest predicted scores for high and low memorability images



(b) Farthest predicted scores for high (top row) and low (bottom row) memorability images

Fig. 6 (a) shows predicted scores where our model performed best. The predicted scores were close for both high and low memorability images. (b) shows cases where our model did not perform well. In the case of high memorability images, the model underpredicted the scores and overpredicted for low memorability images. The green boxes (top right of the images) are the Ground Truth scores and the red boxes (top left of the images) is the Predicted scores.

5 Conclusion

In this study, we introduced a novel multi-modal approach to image memorability prediction by integrating spatial structural information, visual saliency, and semantic information through a GCN-based framework. Our model achieved competitive results across four memorability datasets, with state-of-the-art performance on Object-Mem, SUN, and FIGRIM, showing the effectiveness of combining multi-frequency filter banks, saliency mapping, and image captioning embeddings to capture complex relationships within images. These results highlight the importance of frequency filtering in visual data and the contextual relevance of image regions for enhancing memorability prediction. Using attention mechanisms and advanced graph embeddings, our approach aligns closely with human perception. Future research on image memorability prediction could explore challenging topics such as expanding the model to predict video memorability where temporal dynamics and evolving visual contexts over a sequence of frames influence retention. Another area for improvement is enhancing cross-modal relations in the model.

Declarations

Data availability: The datasets used in this research are all publicly available. The code will be released for academic usage upon acceptance.

Funding: No funding was received for conducting this study.

Competing interests: The authors declare no competing interests.

References

- [1] Isola P, Xiao J, Torralba A, Oliva A. What makes an image memorable? CVPR 2011. 2011;p. 145–152.
- [2] Rust NC, Mehrpour V. Understanding image memorability. Trends in cognitive sciences. 2020;24(7):557–568.
- [3] Bylinskii Z, Isola P, Bainbridge CM, Torralba A, Oliva A. Intrinsic and extrinsic effects on image memorability. Vision Research. 2015;116:165–178.
- [4] Squalli-Houssaini H, Duong NQK, Gwenaelle M, Demarty CH. Deep Learning for Predicting Image Memorability. In: ICASSP; 2018. p. 2371–2375.
- [5] Isola P, Parikh D, Torralba A, Oliva A. Understanding the Intrinsic Memorability of Images. In: NeurIPS; 2011. p. 2429–2437.
- [6] Bainbridge WA, Isola P, Oliva A. The intrinsic memorability of face photographs. Journal of Experimental Psychology: General. 2013 Nov;142(4):1323–1334. <https://doi.org/10.1037/a0033872>.
- [7] Dubey R, Peterson J, Khosla A, Yang MH, Ghanem B. What Makes an Object Memorable? In: ICCV; 2015. p. 1089–1097.

- [8] Khosla A, Xiao J, Torralba A, Oliva A. Memorability of Image Regions. In: NeurIPS. Lake Tahoe, USA; 2012. .
- [9] Kim J, Yoon S, Pavlovic V. Relative Spatial Features for Image Memorability. In: ACM International Conference on Multimedia; 2013. p. 761–764.
- [10] Yoon S, Kim J. Object-Centric Scene Understanding for Image Memorability Prediction. In: MIPR; 2018. p. 305–308.
- [11] Mancas M, Le Meur O. Memorability of natural scenes: The role of attention. In: ICIP; 2013. p. 196–200.
- [12] Fajtl J, Argyriou V, Monekosso D, Remagnino P. AMNet: Memorability Estimation with Attention. In: CVPR; 2018. p. 6363–6372.
- [13] Jing P, Su Y, Nie L, Gu H. Predicting image memorability through adaptive transfer learning from external sources. *IEEE Transactions on Multimedia*. 2016;19(5):1050–1062.
- [14] Baveye Y, Cohendet R, Perreira Da Silva M, Le Callet P. Deep learning for image memorability prediction: The emotional bias. In: ACM international conference on Multimedia; 2016. p. 491–495.
- [15] Zarezadeh S, Rezaeian M, Sadeghi MT. Image memorability prediction using deep features. In: Iranian conference on electrical engineering. IEEE; 2017. p. 2176–2181.
- [16] Banna TT, Deb S, Rahman S, Rahman S. GEMM: A Graph Embedded Model for Memorability Prediction. In: IJCNN. IEEE; 2023. p. 1–8.
- [17] Awh E, Vogel EK, Oh SH. Interactions between attention and working memory. *Neuroscience*. 2006;139(1):201–208.
- [18] Berg D, Itti L. Memory, eye position and computed saliency. *Journal of Vision*. 2008;8(6):1164–1164.
- [19] Foulsham T, Underwood G. What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of vision*. 2008;8(2):6–6.
- [20] Veas EE, Mendez E, Feiner SK, Schmalstieg D. Directing attention and influencing memory with visual saliency modulation. In: Proceedings of SIGCHI; 2011. p. 1471–1480.
- [21] Constant M, Liesefeld HR. Massive effects of saliency on information processing in visual working memory. *Psychological Science*. 2021;32(5):682–691.

- [22] Veldhuis A. The influence of self-relevant information on attention and memory [PhD thesis]. Oxford Brookes University; 2019.
- [23] Borkin MA, Bylinskii Z, Kim NW, Bainbridge CM, Yeh CS, Borkin D, et al. Beyond Memorability: Visualization Recognition and Recall. *IEEE Transactions on Visualization and Computer Graphics*. 2016;22(1):519–528. <https://doi.org/10.1109/TVCG.2015.2467732>.
- [24] Lupyan G, Rakison DH, McClelland JL. Language is not just for talking: Redundant labels facilitate learning of novel categories. *Psychological science*. 2007;18(12):1077–1083.
- [25] Linardos A, Kümmerer M, Press O, Bethge M. DeepGaze IIE: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. In: *ICCV*; 2021. p. 12919–12928.
- [26] Khosla A, Raju AS, Torralba A, Oliva A. Understanding and Predicting Image Memorability at a Large Scale. In: *ICCV*; 2015. p. 2390–2398. ISSN: 2380-7504.
- [27] Isola P, Xiao J, Parikh D, Torralba A, Oliva A. What Makes a Photograph Memorable? *IEEE TPAMI*. 2014;36(7):1469–1482. <https://doi.org/10.1109/TPAMI.2013.200>.
- [28] Oliva A, Torralba A. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*. 2004;42:145–175.
- [29] Lazebnik S, Schmid C, Ponce J. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). vol. 2; 2006. p. 2169–2178.
- [30] Shechtman E, Irani M. Matching Local Self-Similarities across Images and Videos. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition; 2007. p. 1–8.
- [31] Xiao J, Hays J, Ehinger KA, Oliva A, Torralba A. SUN database: Large-scale scene recognition from abbey to zoo. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 2010. p. 3485–3492.
- [32] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol. 1; 2005. p. 886–893 vol. 1.
- [33] Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2010;32(9):1627–1645. <https://doi.org/>

- [34] Khosla A, Xiao J, Isola P, Torralba A, Oliva A. Image Memorability and Visual Inception. In: SIGGRAPH Asia 2012 Technical Briefs; 2012. .
- [35] Celikkale B, Erdem A, Erdem E. Predicting memorability of images using attention-driven spatial pooling and image semantics. *Image and vision Computing*. 2015;42:35–46.
- [36] Rahman S, Bruce ND. Factors underlying inter-observer agreement in gaze patterns: Predictive modelling and analysis. In: *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*; 2016. p. 155–162.
- [37] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet Large Scale Visual Recognition Challenge. *IJCV*. 2015;115(3):211–252. <https://doi.org/10.1007/s11263-015-0816-y>.
- [38] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: *CVPR*; 2016. p. 770–778.
- [39] Perera S, Tal A, Zelnik-Manor L. Is Image Memorability Prediction Solved? In: *CVPR Workshops*; 2019. p. 800–808.
- [40] Itti L. Visual salience. *Scholarpedia*. 2007;2(9):3327.
- [41] Wolfe JM, Horowitz TS. What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*. 2004;5(6):495–501.
- [42] She D, Lai YK, Yi G, Xu K. Hierarchical layout-aware graph convolutional network for unified aesthetics assessment. In: *CVPR*; 2021. p. 8475–8484.
- [43] Ghosal K, Smolic A. Image aesthetics assessment using graph attention network. In: *ICPR*. IEEE; 2022. p. 3160–3167.
- [44] He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. In: *Proceedings of the IEEE international conference on computer vision*; 2017. p. 2961–2969.
- [45] Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph Attention Networks. *ICLR*. 2018;.
- [46] Kipf TN. Semi-supervised classification with graph convolutional networks. In: *International Conference on Learning Representations*; 2017. .
- [47] Rahman S, Rahman S, Shahid O, Abdullah MT, Surov JA. Classifying Eye-Tracking Data Using Saliency Maps. In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE; 2021. p. 9288–9295.

- [48] Judd T, Ehinger K, Durand F, Torralba A. Learning to predict where humans look. In: ICCV; 2009. p. 2106–2113.
- [49] Tseng PH, Carmi R, Cameron IG, Munoz DP, Itti L. Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of vision*. 2009;9(7):4–4.
- [50] Zhang L, Tong MH, Marks TK, Shan H, Cottrell GW. SUN: A Bayesian framework for saliency using natural statistics. *Journal of vision*. 2008;8(7):32–32.
- [51] Fey M, Sunil J, Nitta A, Puri R, Shah M, Stojanović B, et al. PyG 2.0: Scalable Learning on Real World Graphs. In: *Temporal Graph Learning Workshop @ KDD*; 2025. .
- [52] Balcilar M, Renton G, Héroux P, Gaüzère B, Adam S, Honeine P. Analyzing the Expressive Power of Graph Neural Networks in a Spectral Perspective. In: *International Conference on Learning Representations*; 2021. .
- [53] Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. In: *3rd International Conference on Learning Representations (ICLR)*. San Diego, CA, USA; 2015. .
- [54] Kiros R, Salakhutdinov R, Zemel RS. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:14112539*. 2014;.
- [55] Van der Maaten L, Hinton G. Visualizing data using t-SNE. *Journal of machine learning research*. 2008;9(11).