

Optimizing Cross-Modal Image–Text Matching Algorithms Based on Graph Neural Networks: Implications for Textile Applications

Ke Dong¹

¹School of Computer and Artificial Intelligence, Xihang University, Xi'an City 710077, Shaanxi Province, China

Corresponding author: Ke Dong (e-mail: wingsing111@163.com)

ABSTRACT The challenge of image–text matching stems from differences in representation and structural semantics, particularly in complex scenarios such as digital textile and apparel analysis, resulting in a semantic gap that degrades matching performance. This paper proposes an optimized graph neural network (GNN)-based algorithm to improve cross-modal matching accuracy. Images and text are first represented as graph structures, where image regions and textual keywords are modeled as nodes and their relationships are encoded as edges. The GNN enables progressive information propagation and feature fusion across modalities, while an adaptive semantic alignment module dynamically adjusts propagation weights according to node-level semantic similarity. A multi-scale GNN architecture is further introduced to capture hierarchical semantic features and is optimized using a cross-modal contrastive loss that maximizes the similarity of correctly matched image–text pairs. Such structured multimodal semantic modeling also provides methodological insights for intelligent information fusion and heterogeneous data interpretation in advanced electromagnetic sensing and communication systems. Experimental results demonstrate Recall@1 scores of 70.1% and 60.5% on the MSCOCO and Flickr30k datasets, respectively, with a mean rank of 3.1 and mutual information of 0.75. The proposed method effectively narrows the semantic gap and significantly enhances cross-modal matching accuracy, demonstrating substantial potential for intelligent retrieval and analysis in the textile industry.

INDEX TERMS cross-modal matching, graph neural networks, semantic alignment, multimodal information fusion, electromagnetic information processing

I. INTRODUCTION

CROSS-modal image–text alignment is a central challenge in computer vision and natural language processing, gaining importance as the volume of visual and textual data in digital platforms expands. Applications such as automatic image captioning, visual question answering, and content retrieval critically depend on accurate cross-modal matching [1,2]. In the textile and apparel industry, matching product images with textual descriptions—including fabric, style, and craftsmanship details—is essential for e-commerce search engines and supply chain management. Images primarily convey visual information, such as shape, color, and spatial structure, whereas text is based on linguistic symbols, encompassing grammar, semantics, and context. These representational differences create a substantial semantic gap, complicating accurate alignment, especially in dynamic, high-dimensional image content or extended textual descriptions. The misalignment is particularly pronounced in cases involving abstract imagery or complex narrative descriptions. Although deep neural net-

works have advanced cross-modal matching [3,4], existing methods continue to face challenges in semantic alignment and often fall short of performance requirements in real-world applications [5,6]. Recent research has explored the potential of graph neural networks (GNNs), leveraging their ability to model structured relationships. For instance, Dong et al. developed a graph convolutional framework with adversarial mechanisms to enhance feature discriminability and cross-modal stability [7]. Mounika and Yadav utilized dynamic attention within multimodal graphs for fine-grained matching [8], while Cui et al. introduced graph reasoning units to explicitly model entity relationships, improving alignment accuracy [9]. Liu et al. incorporated semantic relationship graphs to facilitate fine-grained interactions between images and text [10]. Although GNNs can effectively model structured relationships, many current approaches do not adequately address semantic drift caused by noisy nodes in traditional GNN-based alignment methods [11,12].

To address semantic alignment challenges, Park et al. proposed a cross-modal semantic alignment module that

substantially improved semantic consistency between images and text [13]. Tao et al. designed a memory-based cross-modal alignment network for medical report generation, effectively managing information matching between lengthy text and images [14]. Sun et al. combined adversarial learning with a Transformer structure to introduce a semantic alignment transformer and added a hashing mechanism to improve retrieval efficiency [15]. Messina et al. adopted a fine-grained image–text alignment strategy using the Transformer encoder to capture multi-level semantics in complex scenarios [16]. Hu et al. investigated cross-modal event localization, reinforcing temporal and semantic associations via semantic collaborative learning [17]. Peng et al. proposed a relationship aggregation mechanism for cross-graph semantic fusion [18], and Shi et al. addressed image–text retrieval with incomplete modalities through deep correlation transfer [19]. Zeng et al. constructed a multi-instance graph structure guided by keywords to diversify image retrieval [20].

In cross-modal retrieval, researchers continue to develop more effective semantic alignment and feature modeling strategies to improve image–text matching [21–23]. Aseniserere and Burnett [24] highlighted challenges in implementing business intelligence systems in Nigeria’s textile industry, such as infrastructural limitations and skill shortages. Bermeo-Giraldo et al. [25] emphasized the growing role of digital transformation and the integration of AI and big data to optimize production processes. Weible et al. [26] stressed the importance of improved testing methods for outdoor apparel, underscoring the need for advanced data utilization—an area where GNNs for image–text matching can provide further optimization. Guo et al. [27] enhanced cross-modal matching robustness through alignment reconstruction mechanisms, increasing adaptability to varying sample distributions.

To bridge the semantic gap in image–text matching, this paper proposes a GNN-based optimization algorithm designed to enhance matching accuracy [28–30]. The innovation lies in mapping images and text to graph structures and employing GNNs for multi-level feature propagation and semantic fusion. Unlike traditional approaches, GNNs more accurately model semantic relationships between image regions and text keywords, and gradual information propagation prevents semantic distortion during feature fusion [31,32]. An adaptive semantic alignment module is introduced to dynamically adjust propagation weights based on semantic similarity, thereby improving alignment accuracy. Additionally, a multi-scale GNN structure captures and fuses semantic features from low to high levels using skip connections, further boosting crossmodal matching accuracy. These innovations facilitate the integration of image and text information in complex scenarios, effectively resolving the semantic gap and substantially enhancing matching accuracy. Commercially, this framework supports applications such as image search and visual question answering, with broad potential in sectors like healthcare and e-commerce.

Specifically, in the textile industry, it promises more accurate fabric databases, intelligent fashion trend analysis, and automated quality control, delivering significant commercial and industrial value.

II. METHODS

GRAPH STRUCTURE MAPPING OF IMAGES AND TEXT CONSTRUCTION OF IMAGE REGION GRAPH

Images are represented structurally by extracting region nodes using a backbone region detection method, and spatial and semantic relationships among regions are encoded as graph edges [33]. Node construction begins with a pre-trained Faster R-CNN model (ResNet-101 backbone) to extract 36 high-confidence target regions per image, based on MS COCO pre-trained weights. Each region is transformed into a 2,048-dimensional feature vector via RoI Pooling, serving as the initial node feature in the graph. Edges between nodes are established according to geometric center distance, relative area ratio, spatial overlap rate (Intersection over Union, IoU), and category similarity. The adjacency matrix $A \in R^{N \times N}$, where $N = 36$, and edge weights are computed as follows (Equation 1):

$$A_{ij} = \exp\left(-\frac{\|c_i - c_j\|^2}{\sigma^2}\right) \cdot \delta(IoU_{ij} > \tau) + \lambda \cdot \text{sim}(f_i, f_j) \quad (1)$$

In Equation (1), c_i and c_j denote the center coordinates of regions i and j ; σ is the standard deviation of the Gaussian kernel; τ is the IoU threshold; $\text{sim}(f_i, f_j)$ denotes region semantic similarity (computed via category labels or feature cosine similarity); and λ controls the influence of semantic similarity on edge weights. The resulting graph is a weighted undirected graph, with each image instance corresponding to a node set V_g and edge set E_g , forming graph $G_g = (V_g, E_g)$. This structure preserves the local visual characteristics of image regions while effectively capturing spatial and semantic connectivity through edge construction. The Faster R-CNN model is initialized with MS COCO pre-trained weights and a ResNet-101-FPN backbone. The number of region proposals is set to 100, with the top 36 highest-confidence regions retained. RoI Pooling output has a 2,048-dimensional feature. All images are resized to 800 pixels on the shorter side (maximum 1,333 pixels on the longer side) and normalized using standard ImageNet procedures.

CONSTRUCTION OF TEXT SEMANTIC GRAPH

Text graph structures are constructed through node segmentation and edge connection, based on syntactic dependency parsing and semantic association within word embeddings [34]. The input text is segmented, and syntactic dependencies are extracted using the Stanford CoreNLP toolkit. Each content word or noun phrase becomes a node; after removing stop words, approximately 10–15 key nodes are retained per text. Node features are generated by the BERT-base

(case-insensitive) model, yielding a 768-dimensional static embedding vector for each node. The model's parameters are fixed during training, serving only as a feature extractor.

Edges comprise two types: syntactic edges, which connect word pairs with explicit syntactic dependencies, and semantic edges, which connect pairs whose embedding cosine similarity exceeds threshold $\theta = 0.7$. In the final adjacency matrix $A_t \in R^{M \times M}$, M is the number of nodes, and the edge weight is determined by the following formula (2):

$$A_{ij} = \begin{cases} 1, & \text{if } (i, j) \in \text{dep}, \\ \cos(e_i, e_j), & \text{if } \cos(e_i, e_j) > \theta, \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

In Equation (2), e_i and e_j are the word embedding representations of nodes i and j , respectively, and dep denotes the set of syntactic dependency edges. Each node also includes part-of-speech (POS) tagging and syntactic roles (e.g., subject, object) as additional features, which are mapped through an embedding layer and concatenated into the node representation. This text semantic graph preserves both structural linguistic relationships (modification, subordination, parallelism) and potential associations among words, enabling the model to capture long-range dependencies and high-level semantic combinations. Figure 1 illustrates the two principal structures: Figure 1(a) depicts an image region graph with the top 12 region nodes (e.g., salient objects and background), positioned in a two-dimensional space. Node size reflects region area, and edges (with thickness proportional to IoU) highlight spatial relationships. Figure 1(b) presents a text graph with 12 semantically related words; nodes are positioned via semantic embedding. Blue solid lines indicate syntactic dependencies, while orange dashed lines denote semantic enhancements, simulating both explicit and implicit connections. This structural approach underpins improved image–text matching accuracy. For clarity, only the top 12 nodes are visualized, though the full model uses 36.

INFORMATION PROPAGATION IN CROSS-MODAL GRAPH NEURAL NETWORKS

INTRA-MODAL INFORMATION PROPAGATION

Following graph construction, both image region and text semantic graphs are processed by a graph neural network to capture internal structural dependencies and contextual information within each modality [35]. The Graph Attention Network (GAT) serves as the propagation framework. At each layer, features of adjacent nodes are aggregated using learnable attention weights. For any node v_i , feature updates are governed by Equation (3):

$$h_i' = \sigma \left(\sum_{j \in N(i)} \alpha_{ij} W h_j \right) \quad (3)$$

Here, h_i is the node's initial feature; $W \in R^{d' \times d}$ is a learnable linear transformation matrix, and α_{ij} is the

attention weight from node j to node i , which is calculated as follows:

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(a^T [W h_i \| W h_j]))}{\sum_{k \in N(i)} \exp(\text{LeakyReLU}(a^T [W h_i \| W h_k]))} \quad (4)$$

In Equation (4), $a \in R^{2d}$ represents a learnable vector; $\|$ denotes vector concatenation, and σ is the activation function (default: ELU). To enhance expressiveness and mitigate oversmoothing (where node representations become homogenized as layers increase), a multi-head attention mechanism is employed. This enables different attention heads to focus on distinct semantic substructures, preserving node-level differentiation.

For image graphs, three GNN layers are applied, reducing node feature dimensions from 2,048 to 512, which is retained in subsequent layers. For text graphs, two layers are used, reducing BERT's 768-dimensional output to 512. After each propagation round, contextual information is disseminated across the graph, enabling local features to capture structural dependencies.

CROSS-MODAL GRAPH ALIGNMENT AND FUSION

After modeling intra-modal features, a dual-graph cross-attention mechanism is introduced to enable information fusion and semantic alignment between image and text graphs. Specifically, each image node serves as a query, while all text nodes act as keys and values in a Transformer-style attention mechanism. Let $V_g = \{v_i\}_{i=1}^n$ be the sets of image and $V_t = \{t_j\}_{j=1}^m$ be the set of text nodes, respectively. Cross-modal attention updates image nodes as follows (Equation 5):

$$\begin{aligned} \hat{v}_i &= \sum_{j=1}^m \beta_{ij} \cdot (W_V t_j), \\ \beta_{ij} &= \frac{\exp((W_Q v_i)^T (W_K t_j) / \sqrt{d})}{\sum_{k=1}^m \exp((W_Q v_i)^T (W_K t_k) / \sqrt{d})} \end{aligned} \quad (5)$$

Matrices $W_Q, W_K, W_V \in R^{d \times d}$ are learnable weights; $d = 512$ is the embedding dimension; and β_{ij} is the attention weight. This operation encodes each image node's semantic relevance to all text nodes, enabling cross-modal feature updates. The process is mirrored for text nodes. To address inconsistent feature distributions, batch normalization and residual connections are employed after cross-graph attention updates, fusing original and updated features (Equation 6):

$$v_i^{out} = \text{BN}(\hat{v}_i + v_i), \quad t_j^{out} = \text{BN}(\hat{t}_j + t_j) \quad (6)$$

The final node features are used for image–text similarity calculation and matching.

A symmetric structure enables bidirectional information propagation, supporting cross-modal attention from image

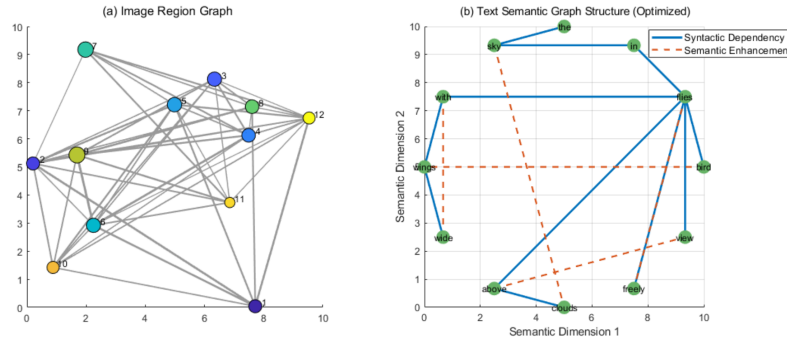


FIGURE 1. Visualization of graph structure in cross-modal image-text matching: (a) image region graph; (b) text semantic graph

to text and vice versa, ensuring continuous information flow and structural consistency for downstream alignment.

During training, GNN propagation and cross-modal attention occur in parallel within each batch. Each image-text pair forms a positive sample, while randomly selected negative samples support contrastive learning and loss optimization. This iterative process gradually reduces the semantic gap, aligning node representations and improving matching accuracy and stability.

Figure 2(a) displays the attention weight matrix for 36 image region nodes; color intensity indicates attention strength, highlighting spatial and semantic relationships. Figure 2(b) presents cross-modal attention between 12 text nodes and 36 image nodes; high attention values (>0.7) reflect strong semantic alignment, enhancing fine-grained matching.

To avoid dominance by the visual modality, the model employs bidirectional normalization and symmetric residual control. Both image and text features are projected to 512 dimensions and L2-normalized. Crossattention uses a symmetric structure with bidirectional interactions, and batch normalization with gated fusion weights ensures balanced fusion based on semantic relevance.

ADAPTIVE SEMANTIC ALIGNMENT MODULE NODE SEMANTIC SIMILARITY MODELING

Effective cross-modal information propagation depends on accurately measuring semantic relevance between image and text nodes. An adaptive alignment mechanism constructs a semantic similarity matrix for node pairs [36]. All nodes are L2-normalized, and similarity scores are calculated via inner product, forming a preliminary similarity matrix (7):

$$S_{ij} = \frac{v_i^T t_j}{\|v_i\| \cdot \|t_j\|} \quad (7)$$

Here, v_i represents the feature vector of the i -th node in the image graph, and t_j represents the feature vector of the j -th node in the text graph. The normalized similarity matrix serves as the cross-modal attention weight and is used as a coefficient to adjust propagation strength between semantically consistent node pairs. This mechanism does not rely on fixed graph edges; instead, it dynamically constructs soft connections based on semantic relevance,

enabling fine-grained alignment. Semantically similar node pairs receive higher propagation weights, while unrelated nodes are suppressed.

INTEGRATION AND UPDATE STRATEGY

After calculating semantic alignment weights, the model performs weighted feature fusion. For each image node, a fused representation is computed based on its attention distribution over all text nodes; text nodes are updated similarly. Residual connections preserve original semantics, and a gating factor controls the fusion ratio.

The final output for each node consists of a fused cross-modal feature and its original feature. During training, gating parameters are generated via a learnable sigmoid function, allowing dynamic adaptation to input complexity.

To enhance contextual extension, a lightweight graph convolution is applied to updated node features post-alignment, with propagation limited to two hops to prevent oversmoothing. Figure 3(a) shows the distribution of gating factors on the MSCOCO validation set. Most values are low, indicating that semantic transfer is suppressed for most nodes, while key nodes receive higher weights. Figure 3(b) illustrates the distribution of Euclidean distances between cross-modal nodes before and after alignment. After alignment, semantically similar nodes are brought closer, and unrelated nodes are further separated, enhancing discriminability and semantic separation.

The alignment process is dynamically computed within each training batch and executed in parallel with GNN propagation, incurring minimal computational overhead (the alignment module comprises less than 5% of total network parameters), ensuring efficiency and scalability.

MULTI-SCALE GRAPH NEURAL NETWORK ARCHITECTURE

HIERARCHICAL GRAPH CONVOLUTIONAL MODEL DESIGN

To model semantic features at varying granularity, a multi-scale GNN architecture is constructed, applying multi-layer graph convolution to capture both local and global semantic information. Image and text graphs undergo three-layer stacked graph convolutional networks (GCNs), employing

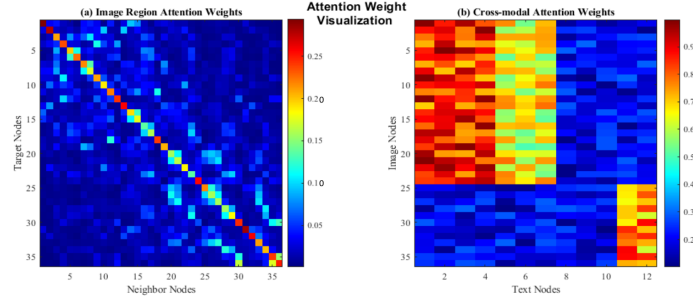


FIGURE 2. Attention weight distribution in multimodal GNNs: (a) image region attention matrix; (b) cross-modal attention matrix

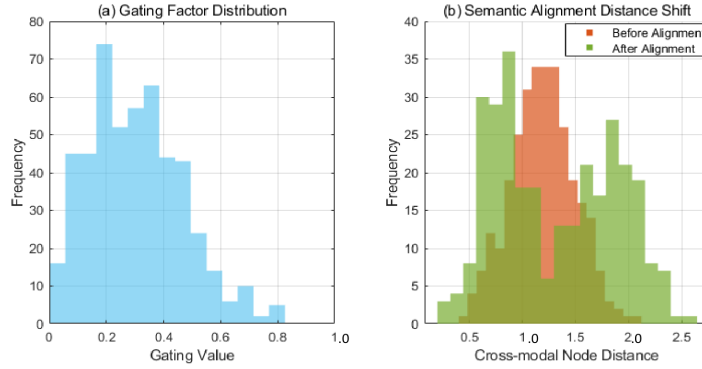


FIGURE 3. Semantic gating and alignment effects: (a) gating factor distribution; (b) changes in semantic alignment distance

different receptive fields and adjacency diffusion strategies at each layer.

The first layer aggregates one-hop neighbors, capturing basic local features. The second layer introduces a residual diffusion matrix for two-hop context and adds feature residuals. The third layer applies high-order convolution with skip connections, integrating information from previous layers and improving the modeling of long-range dependencies.

Each layer is followed by BatchNorm, ReLU, and Dropout (probability 0.3) to prevent overfitting. Node feature dimensions are reduced from 512 to 256, then held constant. Shared parameters are initialized using the Xavier method.

MULTI-SCALE SEMANTIC FUSION STRATEGY

The fusion module concatenates node features from all three convolutional layers along the channel dimension, forming a high-dimensional joint representation. A learnable gating mechanism applies adaptive weighting, dynamically adjusting the contribution of each layer's semantic information. Dimensionality is then reduced to 128 via a fully connected layer and L2 normalization for similarity calculation. For a node in the image graph, the final fused feature is as follows (Equation 8):

$$v_i^{fused} = \sum_{l=1}^L \alpha^{(l)} \cdot v_i^{(l)} \quad (8)$$

Here, $v_i^{(l)}$ denotes the node's representation after the l -th convolution, and $\alpha^{(l)} \in [0, 1]$ is the learnable fusion

coefficient for each layer, with $\sum_{l=1}^L \alpha^{(l)} = 1$ (coefficients derived via softmax). The text graph uses an identical structure to ensure consistency and prevent information imbalance. Fused node features are projected through a non-linear layer, compressed to 128 dimensions, and normalized for downstream similarity matching.

Figure 4 presents the architecture, with input layers receiving node features and adjacency matrices. The first layer captures local features (256-dimensional output), the second enhances context with residual diffusion, and the third integrates information through skip connections and high-order convolution. The fusion module concatenates outputs, applies adaptive weighting, reduces dimensionality, and normalizes the resulting embeddings.

CROSS-MODAL CONTRASTIVE LOSS OPTIMIZATION GLOBAL EMBEDDING CONSTRUCTION AND SIMILARITY CALCULATION

To enable precise matching, node-level features are aggregated into a global embedding vector by weighted averaging or max pooling, yielding a 128-dimensional representation for each image and text. All embeddings are L2-normalized to facilitate comparison in a common space, mitigating amplitude differences and enhancing cross-modal comparability. Embeddings are computed in mini-batches (batch size 128), and a similarity matrix is constructed for all image-text pairs.

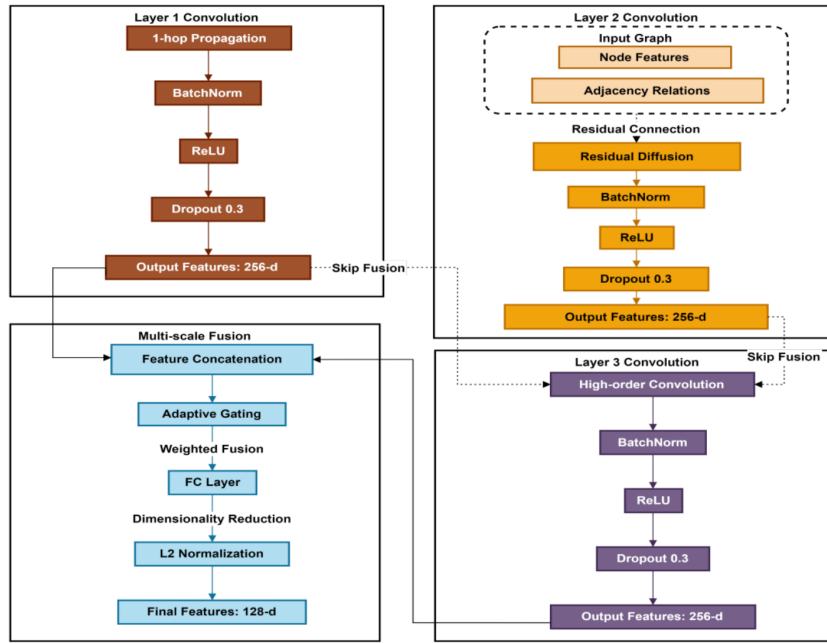


FIGURE 4. Core architecture of the multi-scale GNN: three-layer graph convolutional module and fusion module

CONTRASTIVE LOSS TRAINING MECHANISM

A bidirectional contrastive loss function is employed, optimizing both image-to-text and text-to-image alignment. This approach maximizes similarity for correct matches and minimizes it for mismatched pairs, effectively addressing the semantic gap.

A temperature parameter (0.07) smooths the similarity distribution, preventing local minima and ensuring robust gradient flow. Batch normalization stabilizes training, and regularization reduces overfitting, improving generalization.

Figure 5(a) shows rapid convergence of training and validation losses across epochs, with notable changes corresponding to learning rate adjustments. Figure 5(b) visualizes the embedding space after training. Blue dots represent image features, green dots are semantically aligned text features, and red dots denote mismatched text features. Aligned pairs cluster together, while mismatched samples are peripheral, demonstrating effective cross-modal alignment and discrimination.

III. EFFECTIVENESS EVALUATION

This study employs multiple established cross-modal benchmarks, with MSCOCO and Flickr30k as primary datasets. MSCOCO includes 123,000 annotated images (each with five textual descriptions), while Flickr30k contains 30,000 images (each with five sentences). The experiments strictly follow community protocols: MSCOCO uses the official 118,287/5,000/5,000 (train/validation/test) split, and Flickr30k follows the standard 29,000/1,000/1,000 split. The 7:1:2 ratio mentioned is approximate; experiments replicate Karpathy et al.'s 1K test protocol, ensuring comparability and reproducibility.

RECALL

Recall measures the proportion of true matches detected by the model, a critical metric in retrieval and recommendation scenarios. It reflects the model's sensitivity and completeness in semantic alignment. Table 1 compares recall across methods for image-text retrieval on MSCOCO and Flickr30k using R@1, R@5, and R@10. The proposed method achieves the highest performance on all metrics for both datasets. On MSCOCO, R@1 reaches 70.1%, an 18.1-point improvement over VSE++. R@10 is 96.3%, surpassing AMFM. On Flickr30k, R@1 is 60.5%, outperforming SCAN and DualPath, demonstrating superior cross-modal alignment.

PRECISION

Precision measures the proportion of true matches among all retrieved samples. All metrics follow Karpathy's 1K test protocol, with R@k representing the proportion of queries hitting at least one positive sample, and P@k representing the average proportion of correct matches in the top-k.

Table 2 presents precision results for MSCOCO and Flickr30k using P@1, P@5, and P@10. Precision declines as the ranking range increases, consistent with its definition. The proposed method outperforms all baselines, achieving 66.8% P@1 on MSCOCO (a 16.7-point improvement over VSE++) and 56.0% P@1 on Flickr30k, exceeding all comparators.

MEAN RANK

Mean rank evaluates the model's ability to correctly rank matching pairs. For each query, all candidates are ranked by matching score, and the average rank of the correct match is computed. Lower values indicate better performance.

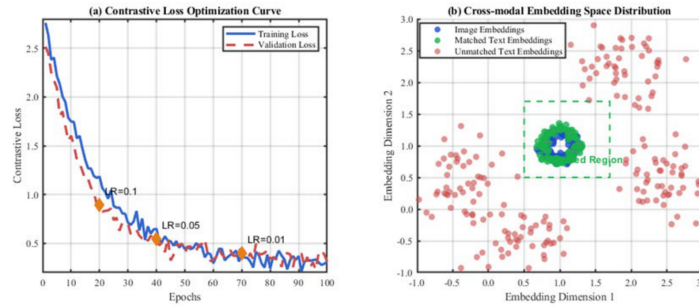


FIGURE 5. Visual analysis of contrastive learning: (a) training and validation loss curves; (b) cross-modal embedding space distribution. Orange dots represent different learning rates during iterations, while red dots represent mismatched text features

TABLE 1. Recall of different methods on MSCOCO and Flickr30K datasets

Method	MSCOCO R@1	MSCOCO R@5	MSCOCO R@10	Flickr30k R@1	Flickr30k R@5	Flickr30k R@10
(%)	(%)	(%)	(%)	(%)	(%)	
VSE++	52.0	80.5	89.4	41.3	70.2	81.1
SCAN	55.2	82.1	90.0	45.8	74.4	84.5
BFAN	59.3	84.8	92.3	48.6	76.2	85.6
CM-GNN	60.4	85.2	92.6	50.1	78.3	87.0
DualPath	61.7	86.0	93.1	51.4	79.0	87.8
SGRAF	63.5	87.1	94.0	53.2	80.2	88.6
CAMP	64.1	87.6	94.3	54.0	81.1	89.2
MTFN	66.0	88.4	94.8	56.3	82.7	90.1
AMFM	67.5	89.3	95.1	57.6	83.6	90.9
Proposed Method	70.1	91.0	96.3	60.5	85.2	92.0

TABLE 2. Precision of different methods on MSCOCO and Flickr30K datasets

Method	MSCOCO P@1	MSCOCO P@5	MSCOCO	Flickr30k P@1	Flickr30k P@5	Flickr30k P@10
(%)	(%)	P@10 (%)	(%)	(%)	(%)	
VSE++	50.1	48.3	42.7	39.6	37.2	33.8
SCAN	53.4	50.5	44.2	43.2	39.8	35.7
BFAN	56.7	52.9	46.3	45.6	41.5	37.0
CM-GNN	57.9	54.1	47.2	47.0	42.7	38.4
DualPath	59.0	55.0	48.1	48.1	43.6	39.1
SGRAF	60.7	56.8	49.3	49.8	45.0	40.5
CAMP	61.2	57.5	49.8	50.3	45.4	40.9
MTFN	63.0	59.1	51.0	52.1	46.8	42.3
AMFM	64.4	60.4	52.3	53.4	47.6	43.1
Proposed Method	66.8	62.7	54.0	56.0	49.8	45.0

Figure 6(a) shows that the proposed method achieves the best average rank at 3.1, indicating robust retrieval accuracy and stability. Figure 6(b) illustrates that 42% of correct matches are ranked first, with decreasing proportions at lower ranks, underscoring the model's strong retrieval performance.

MUTUAL INFORMATION

Mutual information quantifies the degree of information shared between image and text embeddings. Higher values indicate better cross-modal semantic alignment. Joint and marginal probabilities are estimated from embedding distributions, and information gain is calculated to assess alignment depth and breadth. Table 3 compares mutual information across methods. The proposed method achieves 0.75, outperforming all others. Standard deviations are low, indicating stable results. These findings underscore the importance of mutual information for cross-modal retrieval.

TABLE 3. Comparison of mutual information of different methods

Method	Mutual Information	Std Dev
VSE++	0.42	0.020
SCAN	0.48	0.015
BFAN	0.52	0.018
CM-GNN	0.55	0.016
DualPath	0.58	0.014
SGRAF	0.61	0.013
CAMP	0.63	0.012
MTFN	0.65	0.011
AMFM	0.67	0.010
Proposed Method	0.75	0.009

IV. DISCUSSION

This study demonstrates the superior capability of the GNN framework to address the cross-modal semantic gap, validated on MSCOCO and Flickr30k. Although not yet tested on textile-specific datasets, its strong semantic alignment and structured reasoning are highly relevant for textile and

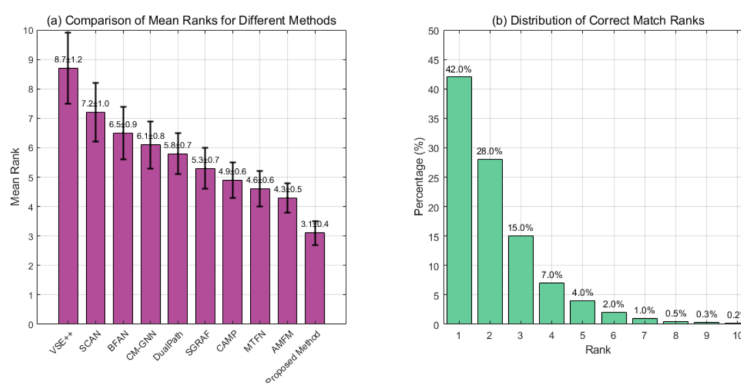


FIGURE 6. Mean rank comparison: (a) average ranking; (b) distribution of correct match rankings

fashion informatics [37]. Textile and fashion data present unique complexities. Apparel images contain hierarchical details—garment silhouette, local components (e.g., collars, sleeves), fabric textures, and patterns—while text descriptions use specialized terminology and subjective language (e.g., draped silk fabric, Bohemian-style paisley print). The proposed model addresses these challenges effectively: the image region graph decomposes garments into structured nodes, enabling precise relationships between components, crucial for distinguishing subtle style differences [38]. The text semantic graph goes beyond keyword matching, capturing complex relationships such as slim fit and A-line silhouette via syntactic and semantic dependencies [39].

The multi-scale GNN architecture is well-suited for modeling fashion semantics, combining local features (e.g., fabric texture, seams) at shallow layers with global style attributes (e.g., professional, street, bohemian) at deeper layers [40]. The adaptive semantic alignment module accurately links textual terms (e.g., jacquard) to complex weaving patterns in images, enabling fine-grained image-text matching [41]. Such capabilities are foundational for next-generation smart textile applications, including:

- **Intelligent retrieval and personalized recommendation:** Users can query in natural language (e.g., “find a retro-style square-neck puff-sleeve dress”), and the model accurately matches products, supporting personalized recommendations [42].
- **Fashion trend analysis and prediction:** The model analyzes social media and fashion blogs, associating visual designs with emerging keywords to quantify and predict trends [43].
- **Automated cataloging and supply chain management:** The model generates structured descriptions for new garments, improves inventory management, and enables automated quality control [44].

In summary, the model’s strong performance on benchmark datasets provides a foundation for textile and fashion industry applications. Future work will focus on fine-tuning with large-scale, fine-grained fashion datasets to further advance intelligent textile industry transformation.

V. CONCLUSION

This study addresses the semantic gap in cross-modal image-text alignment, a critical issue in textile informatics. A GNN-based enhancement strategy is proposed, mapping visual and linguistic features to structured graphs that capture multi-granularity semantics and inter-modal associations. The multi-scale architecture models micro (e.g., fabric texture) and macro (e.g., overall style) features, while text semantic graphs parse relationships such as “high-count cotton” and “twill denim.” The adaptive alignment module precisely links specific descriptions (e.g., puff sleeves) to corresponding image regions, enabling accurate fabric retrieval, style comparison, and defect detection. Experimental results show Recall@1 scores of 70.1% and 60.5% on MSCOCO and Flickr30k, mean rank of 3.1, and mutual information of 0.75, outperforming nine state-of-the-art baselines. Nonetheless, challenges remain in parsing abstract text and addressing computational cost constraints for real-time applications.

AUTHOR CONTRIBUTIONS

Ke Dong designed, collected and analyzed the data, and drafted the manuscript. Ke Dong conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Ke Dong participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

- [1] Y. Cheng, X. Zhu, J. Qian, F. Wen, and P. Liu, "Cross-modal graph matching network for image-text retrieval," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 18, no. 4, pp. 1-23, 2022, doi: 10.1145/3499027.
- [2] S. Qian, D. Xue, Q. Fang, and C. Xu, "Integrating multi-label contrastive learning with dual adversarial graph neural networks for cross-modal retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4794-4811, 2022, doi: 10.1109/TPAMI.2022.3188547.
- [3] M. Hou, Z. Zhang, C. Liu, and G. Lu, "Semantic alignment network for multi-modal emotion recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 9, pp. 5318-5329, 2023, doi: 10.1109/TCSVT.2023.3247822.
- [4] R. Raman, P. Das, R. Aggarwal, R. Buch, B. Palanisamy, and T. Basant, "Circular Economy Transitions in Textile, Apparel, and Fashion: AI-Based Topic Modeling and Sustainable Development Goals Mapping," *Sustainability*, vol. 17, no. 12, pp. 5342, 2025, doi: 10.3390/su17125342.
- [5] N. Kaur, S. Pandey, and N. Kalra, "Fashion cloth image categorization and retrieval with enhanced intensity using SURF and CNN approach," *International Journal of Clothing Science and Technology*, vol. 37, no. 2, pp. 222-241, 2025, doi: 10.1108/IJCST-03-2024-0074.
- [6] T. Yao, Y. Li, Y. Li, G. Wang, and J. Yue, "Cross-modal semantically augmented network for image-text matching," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 4, pp. 1-18, 2023, doi: 10.1145/3631356.
- [7] X. Dong, L. Liu, L. Zhu, L. Nie, and H. Zhang, "Adversarial graph convolutional network for cross-modal retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1634-1645, 2021, doi: 10.1109/TCSVT.2021.3075242.
- [8] K. Mounika and B. V. R. Yadav, "ATTENTION-GUIDED MULTI-MODAL GRAPH NETWORKS FOR CROSS-MODAL IMAGE-TEXT MATCHING," *Cuestiones de Fisioterapia*, vol. 54, no. 5, pp. 358-372, 2025, doi: 10.48047/69h5ek37.
- [9] Z. Cui, Y. Hu, Y. Sun, J. Gao, and B. Yin, "Cross-modal alignment with graph reasoning for image-text retrieval," *Multimedia Tools and Applications*, vol. 81, no. 17, pp. 23615-23632, 2022, doi: 10.1007/s11042-022-12444-8.
- [10] X. Liu, Y. He, M. Cheung Y, X. Xu, and N. Wang, "Learning relationship-enhanced semantic graph for fine-grained image-text matching," *IEEE transactions on cybernetics*, vol. 54, no. 2, pp. 948-961, 2022, doi: 10.1109/TCYB.2022.3179020.
- [11] J. Pei, K. Zhong, Z. Yu, L. Wang, and K. Lakshmana, "Scene graph semantic inference for image and text matching," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 5, pp. 1-23, 2023, doi: 10.1145/3563390.
- [12] J. Cao, X. Qin, S. Zhao, and J. Shen, "Bilateral cross-modality graph matching attention for feature fusion in visual question answering," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 3, pp. 4160-4171, 2022, doi: 10.1109/TNNLS.2021.3135655.
- [13] P. Park, S. Jang, Y. Cho, and Y. Kim, "SAM: Cross-modal semantic alignments module for image-text retrieval," *Multimedia Tools and Applications*, vol. 83, no. 4, pp. 12363-12377, 2024, doi: 10.1007/s11042-023-15798-9.
- [14] Y. Tao, L. Ma, J. Yu, and H. Zhang, "Memory-based cross-modal semantic alignment network for radiology report generation," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 7, pp. 4145-4156, 2024, doi: 10.1109/JBHI.2024.3393018.
- [15] Y. Sun, M. Wang, and Y. Ma, "Semantic-alignment transformer and adversary hashing for cross-modal retrieval," *Applied Intelligence*, vol. 54, no. 17, pp. 7581-7602, 2024, doi: 10.1007/s10489-024-05501-2.
- [16] N. Messina, G. Amato, A. Esuli, F. Falchi, C. Gennaro, and S. Marchand-Maillet, "Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 17, no. 4, pp. 1-23, 2021, doi: 10.1145/3451390.
- [17] Y. Hu, K. Wang, M. Liu, H. Tang, and L. Nie, "Semantic collaborative learning for cross-modal moment localization," *ACM Transactions on Information Systems*, vol. 42, no. 2, pp. 1-26, 2023, doi: 10.1145/3620669.
- [18] J. Peng S, Y. He, X. Liu, Y. Cheung, X. Xu, and Z. Cui, "Relation-aggregated cross-graph correlation learning for finegrained image-text retrieval," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 2, pp. 2194-2207, 2022, doi: 10.1109/TNNLS.2022.3188569.
- [19] D. Shi, L. Zhu, J. Li, G. Dong, and H. Zhang, "Incomplete cross-modal retrieval with deep correlation transfer," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 5, pp. 1-21, 2024, doi: 10.1145/3637442.
- [20] Y. Zeng, Y. Wang, D. Liao, G. Li, W. Huang, J. Xu, et al., "Keyword-based diverse image retrieval with variational multiple instance graph," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 12, pp. 10528-10537, 2022, doi: 10.1109/TNNLS.2022.3168431.
- [21] Z. Li, H. Lu, H. Fu, F. Meng, and G. Gu, "Csan: Cross-coupled semantic adversarial network for cross-modal retrieval," *Artificial Intelligence Review*, vol. 58, no. 5, pp. 1-27, 2025, doi: 10.1007/s10462-025-11152-7.
- [22] X. Liang, E. Yang, C. Deng, F. Meng, and G. Gu, "CrossFormer: Cross-modal Representation Learning via Heterogeneous Graph Transformer," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 12, pp. 1-21, 2024, doi: 10.1145/3688801.
- [23] Z. Liu, G. Hao, F. Li, X. He, and Y. Zhang, "Multi-modal Sentiment Classification Based On Graph Neural Network And Multi-head Cross-attention Mechanism For Education Emotion Analysis," *Journal of Applied Science and Engineering*, vol. 28, no. 6, pp. 1185-1193, 2025, doi: 10.6180/jase.202506_28(6).0002.
- [24] O. Aseniserare and S. Burnett, "Exploring Challenges and Adoption Factors of Business Intelligence Systems in Nigerian Textile and Apparel Industry," *African Journal of Management and Business Research*, vol. 19, no. 1, pp. 456-488, 2025, doi: 10.62154/ajmbr.2025.019.01033.
- [25] C. Bermeo-Giraldo M, A. Valencia-Arias, M. Rojas E, and S. Cardona-Acevedo, "Research agenda on the evolution of digital transformation in the textile sector: a bibliometric analysis and research trends," *Discover Sustainability*, vol. 6, no. 1, pp. 294, 2025, doi: 10.1007/s43621-025-01091-2.
- [26] L. Weible, T. Domina, R. Dankwa, and L. Agnew, "Recommended base layer ensemble for testing winter outdoor apparel using an infant thermal manikin," *International Journal of Fashion Design, Technology and Education*, pp. 1-11, 2025, doi: 10.1080/17543266.2025.2476543.
- [27] J. Guo and J. Ding, "Robust cross-modal retrieval with alignment refurbishment," *Frontiers of Information Technology & Electronic Engineering*, vol. 24, no. 10, pp. 1403-1415, 2023, doi: 10.1631/FITEE.2200514.
- [28] N. Han, J. Chen, H. Zhang, H. Wang, and H. Chen, "Adversarial multi-grained embedding network for cross-modal text-video retrieval," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 18, no. 2, pp. 1-23, 2022, doi: 10.1145/3483381.
- [29] Y. Cao, "English text classification model based on graph neural networks and contrastive learning," *International Journal of Information and Communication Technology*, vol. 26, no. 25, pp. 48-69, 2025, doi: 10.1504/IJICT.2025.147141.
- [30] Y. Gong, G. Cosma, and A. Finke, "VITR: Augmenting vision transformers with relation-focused learning for cross-modal information retrieval," *ACM Transactions on Knowledge Discovery from Data*, vol. 18, no. 9, pp. 1-21, 2024, doi: 10.1145/3686805.
- [31] X. Qin, L. Li, and G. Pang, "Multi-scale motivated neural network for image-text matching," *Multimedia Tools and Applications*, vol. 83, no. 2, pp. -4407, 2024, doi: 10.1007/s11042-023-15321-0.
- [32] A. Shin, M. Ishii, and T. Narihira, "Perspectives and prospects on transformer architecture for cross-modal tasks with language and vision," *International journal of computer vision*, vol. 130, no. 2, pp. 435-454, 2022, doi: 10.1007/s11263-021-01547-8.
- [33] J. Wang, J. Ke, H. H. Shuai, Y. H. Li, and W. H. Cheng, "Referring expression comprehension via enhanced cross-modal graph attention networks," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 2, pp. 1-21, 2023, doi: 10.1145/3548688.
- [34] X. Dong, X. Zhan, Y. Wei, X. Wei, Y. Wang, M. Lu, et al., "Entity-graph enhanced cross-modal pretraining for instance-level product retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 13117-13133, 2023, doi: 10.1109/TPAMI.2023.3291237.
- [35] S. Xiong, L. Pan, X. Ma, Q. Hu, and E. Beckman, "Unsupervised deep hashing with multiple similarity preservation for cross-modal image-text retrieval," *International Journal of Machine Learning and Cybernetics*, vol. 15, no. 10, pp. -4434, 2024, doi: 10.1007/s13042-024-02154-y.
- [36] Y. Chen, C. Hu, T. Kimura, Q. Li, S. Liu, F. Wu, et al., "Semicmt: Contrastive cross-modal knowledge transfer for iot sensing with semi-paired multi-modal signals," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 4, pp. 1-30, 2024, doi: 10.1145/3699779.
- [37] H. Reddy A, V. Tirunagari, and V. Rajesh K K, "Analytical Prospective on Fashion Industry: Data-Driven Strategies[M]//Predictive Analytics and Generative AI for Data-Driven Marketing Strategies," *Chapman and Hall/CRC*, pp. 33-52, 2024.

- [38] N. Rostamzadeh, S. Hosseini, T. Boquet, W. Stokowiec, Y. Zhang, C. Jauvin, et al., "Fashion-gen: The generative fashion dataset and challenge," arXiv preprint, pp. Jun 21, 2018, doi: 10.48550/arXiv.1806.08317.
- [39] Z. Liu, S. Yan, P. Luo, X. Wang, and X. Tang, "Fashion landmark detection in the wild," in Leibe, B., Matas, J., Sebe, N., Welling, M. In European Conference on Computer Vision 2016. Cham, Germany: Springer International Publishing. 2016, pp. 229-245, doi: 10.1007/978-3-319-46475-6_15.
- [40] B. Pan, J. Xiang, N. Zhang, and R. Pan, "Two-stage attribute-guided dual attention network for fine-grained fashion retrieval," Computer Vision and Image Understanding, Art. no. 104497, 2025, doi: 10.1016/j.cviu.2025.104497.
- [41] S. Bell and K. Bala, "Learning visual similarity for product design with convolutional neural networks," ACM transactions on graphics (TOG), vol. 34, no. 4, pp. 1-10, 2015, doi: 10.1145/2766959.
- [42] W. Yu, X. He, J. Pei, X. Chen, L. Xiong, J. Liu, et al., "Visually aware recommendation with aesthetic features," The VLDB Journal, vol. 30, no. 4, pp. 495-513, 2021, doi: 10.1007/s00778-021-00651-y.
- [43] Z. Al-Halah, R. Stiefelhagen, and K. Grauman, "Fashion forward: Forecasting visual style in fashion," In Proceedings of the IEEE international conference on computer vision. Venice, Italy: IEEE, pp. . p. 388-397, 2017, doi: 10.1109/ICCV.2017.50.
- [44] A. Kovashka, D. Parikh, and K. Grauman, "Whittlesearch: Image search with relative attribute feedback," In 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI, USA: IEEE, pp. . p. 2973-2980, 2012, doi: 10.1109/CVPR.2012.6248026.