VITR: Augmenting Vision Transformers with Relation-Focused Learning for Cross-modal Information Retrieval

YAN GONG, Department of Computing and Informatics, Bournemouth University, Bournemouth, UK
GEORGINA COSMA, Department of Computer Science, Loughborough University, Loughborough, UK
AXEL FINKE, Department of Mathematical Sciences, Loughborough University, Loughborough, UK

The relations expressed in user queries are vital for cross-modal information retrieval. Relation-focused cross-modal retrieval aims to retrieve information that corresponds to these relations, enabling effective retrieval across different modalities. Pre-trained networks, such as Contrastive Language-Image Pre-training networks, have gained significant attention and acclaim for their exceptional performance in various cross-modal learning tasks. However, the Vision Transformer (ViT) used in these networks is limited in its ability to focus on image region relations. Specifically, ViT is trained to match images with relevant descriptions at the global level, without considering the alignment between image regions and descriptions. This article introduces VITR, a novel network that enhances ViT by extracting and reasoning about image region relations based on a local encoder. VITR is comprised of two key components. Firstly, it extends the capabilities of ViT-based cross-modal networks by enabling them to extract and reason with region relations present in images. Secondly, VITR incorporates a fusion module that combines the reasoned results with global knowledge to predict similarity scores between images and descriptions. The proposed VITR network was evaluated through experiments on the tasks of relation-focused cross-modal information retrieval. The results derived from the analysis of the Flickr30K, MS-COCO, RefCOCOg, and CLEVR datasets demonstrated that the proposed VITR network consistently outperforms state-of-the-art networks in image-to-text and text-to-image retrieval.

CCS Concepts: • **General and reference** → **Cross-modal retrieval**; *Multimedia and multimodal retrieval*; • **Computing methodologies** → Reasoning; Machine learning approaches;

Additional Key Words and Phrases: information retrieval, relation-focused cross-modal, visual-semantic embedding, fusion

ACM Reference format:
Yan Gong, Georgina Cosma, and Axel Finke. 2024. VITR: Augmenting Vision Transformers with Relation-Focused Learning for Cross-modal Information Retrieval. *ACM Trans. Knowl. Discov. Data.* 18, 9, Article 220 (October 2024), 21 pages.
https://doi.org/10.1145/3686805

Authors' Contact Information: Yan Gong, Department of Computing and Informatics, Bournemouth University, Bournemouth, UK; e-mail: gongyan1920@163.com; Georgina Cosma (corresponding author), Department of Computer Science, Loughborough University, Loughborough, UK; e-mail: g.cosma@lboro.ac.uk; Axel Finke, Department of Mathematical Sciences, Loughborough University, Loughborough, UK; e-mail: a.finke@lboro.ac.uk.

ACM 1556-472X/2024/10-ART220
https://doi.org/10.1145/3686805

1 Introduction

Due to the escalation of multi-modal multimedia data [43, 52], cross-modal information retrieval has gained significant prominence. Relation-focused cross-modal information retrieval focuses on extracting information that aligns with the relations expressed in user queries, and it is particularly relevant for the development of next-generation search engines. Such capability will result in improved retrieval and ranking performance since the results will be more relevant to the user's query than when relations are not considered. Consider, for example, Figure 1, which shows a description query containing relations, such as 'person holding food'. A system that considers relations of image regions will rank images (e.g., Figure 1(a)) featuring a person holding food as more similar to the query than images (e.g., Figure 1(b)) depicting people and food separately.

Current works use **Visual-Semantic Embedding (VSE)** networks to embed image–description pairs in a shared latent space and calculate similarity scores for retrieval tasks [15]. Pre-trained VSE networks have recently gained popularity in various cross-modal tasks [5, 6, 30, 47], with the **Contrastive Language-Image Pre-Training (CLIP)** network [36] achieving state-of-the-art performance in cross-modal information retrieval. CLIP employs a pre-trained **Vision Transformer (ViT)** and a transformer-based text encoder to encode images and descriptions into a shared embedding space. ViTs use the self-attention mechanism from transformers, allowing the model to capture long-range dependencies and intricate patterns in the input data, resulting in a rich contextual understanding of the visual modality and improved cross-modal understanding [34].

ViTs have been extensively studied for cross-modal information retrieval, but there is still room for improvement, particularly in relation-focused tasks. ViTs divide images into small blocks [29], which can result in a loss of local information compared to **Convolutional Neural Networks (CNNs)** [44]. This limitation becomes apparent when applying ViT-based pre-trained VSE networks to relation-focused tasks, as the models exhibit weak local perception abilities for images and have limited capacity to align image regions with corresponding descriptions. Additionally, the ViT used in contrastive learning [36] connects with a convolutional layer, its primary design objective is to capture global image features rather than aligning image regions with corresponding descriptions [12, 34]. Modifying the internal structure of the transformer to enhance ViTs ability for capturing local image information may result in potential drawbacks in cross-modal tasks, such as the risk of losing global context.

To address the limitations of ViTs in relation-focused cross-modal information retrieval tasks, this article proposes a novel network named **ViT-Relation-Focus (VITR)**. VITR provides relational reasoning of image regions that are extracted by a local encoder and fuses these relations into the pre-trained ViT for relation-focused cross-modal information retrieval tasks. In this article, relational reasoning involves extracting relevant relations between image regions and generating relation-focused local representations of the image to improve cross-modal information retrieval performance. The contributions of this article are as follows:

- The proposed VITR network extends the capabilities of pre-trained networks for focusing on the relationships between image regions and their corresponding descriptions. VITR incorporates a ViT encoder and a text encoder to obtain pre-trained global representations of images and descriptions, along with a CNN-based local encoder for capturing local representations of images. VITR employs a fusion module that integrates the global and local representations of images and descriptions to predict the similarity scores of image and description pairs. While the foundational concepts of relation-focus are rooted in the vision-language domain, VITR significantly extends and innovates upon these ideas by introducing the fusion module. The fusion module enhances the ViT's ability to capture and reason about image region relations based on fused local and global information.

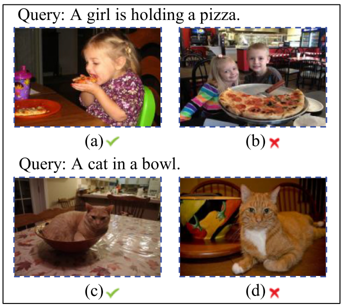


Fig. 1. A retrieval system that considers region relations will rank the image (a) of 'a person holding food' as more relevant to the query description than the image (b) of just 'people and food'.

- VITR leverages a new type of relational reasoning module that first models an image's regions and their relations using a relational graph, then generates local representations aligned with the image's description. Incorporating relation-focused local image representations into VITR improves cross-modal information retrieval performance.
- Inspired by the pre-ranking-reranking strategy, this article develops a module for VITR, named turbo, which selects the top N relevant candidates for the query and sends the necessary candidate embeddings or global representations to relevant modules for further ranking finalisation. The turbo module results in reduced computation time for VITR.
- Extensive experiments were conducted to evaluate VITR on the benchmark datasets Flickr30K and MS-COCO, as well as on RefCOCOg and CLEVR, which involve relation-focused descriptions. VITR outperformed various other state-of-the-art networks, namely CLIP, VSRN++, and VSE∞, in both image-to-text and text-to-image cross-modal information retrieval tasks.

The rest of the article is organised as follows: Section 2 summarises the related work, Section 3 elaborates on the proposed VITR, Section 4 demonstrates the experimental results, Section 5 presents the visualisation results, and Section 6 concludes our work.

2 Related Work

2.1 VSEs

Faghri et al. [11] unveiled VSE++, an elevated VSE architecture that incorporates a fully connected neural network to generate the representations of image features extracted by a faster R-CNN [37] and a **Gated Recurrent Unit (GRU)** network [7] to generate the representations of descriptions. Wang et al. [41] introduced a rare-aware attention network, which aims to address the long-tail effect in image and text matching by exploring and exploiting rare textual content. Lee et al. [19] introduced an attention network designed to unveil the complete latent alignments between image regions and their respective descriptive words. Li et al. [24] introduced the **Visual Semantic Reasoning Network (VSRN)**, designed to augment image features using image region relationships, these relationships being extracted via a **Graph Convolutional Network (GCN)** [49]. Later, Li et al. [25] improved the VSRN by upgrading it to VSRN++, which replaces the word2vector embeddings with pre-trained BERT [9] embeddings. Chen et al. [2] proposed a variant of the VSE network, VSE∞, which leverages a generalised pooling operator to discern the most effective strategy for pooling the representations of images and descriptions. Diao et al. [10] introduced a **Similarity Graph Reasoning (SGR)** network for constructing and reasoning with a graph from attention results, and the Similarity Attention Filtration network for isolating crucial information within these results.

2.2 Pre-Trained Networks for VSEs

The development of pre-trained networks for cross-modal information retrieval has progressed significantly in recent years [5, 6, 13, 30, 47, 51]. Chen et al. [5] presented UNITER, a model that serves as a universal image-text bridge, meticulously pre-trained on four distinct image-text datasets. This network accommodates a diverse array of vision-and-language tasks, generating joint multimodal embeddings through four dedicated pre-training tasks. Yu et al. [47] put forth a methodology that leverages structured knowledge from scene graphs to boost joint representation learning in tasks that intersect vision and language. Lu et al. [30] proposed a novel collaborative two-stream vision-language pre-training approach for image-text retrieval that enhances cross-modal interaction through instance-level alignment, token-level interaction, and task-level interaction. Recently, Radford et al. [36] proposed the pre-trained CLIP which applies contrastive learning to align the global visual representations and textual representations from a dataset including 400 million image–description pairs. The architecture of CLIP involves (1) a text encoder which aims to embed the description as a dimension-reduced representation; (2) an image encoder, commonly using ViT, which aims to embed the image as a representation with the same dimension as the description representation. CLIP has been applied in many tasks recently, such as e-commerce image retrieval [31], video-text retrieval [32], and text-image generation [38]. However, the pre-trained networks, especially CLIP, still lack the ability to effectively match local information in images to their descriptions in cross-modal information retrieval tasks.

2.3 State-of-the-Art Two-Stage Pre-Trained VSE Networks

Li et al. [23] introduced a contrastive loss to **Align the Image and Description Representations Before Fusing (ALBEF)** them through cross-modal attention, which creates coherent and grounded multi-modal representations. Li et al. [26] proposed a **Multi-Level Semantic Alignment Network for Vision-Language Pre-Training (MVPTR)** that enhances semantic alignment across multiple levels by incorporating high-level concepts and a two-stage learning framework, improving performance on image-text retrieval tasks. Zhang et al. [48] presented a **Refined Vision-Language Modeling (RVLM)** network that leverages a homonym sentence rewriting algorithm for token-level supervision and refined contrastive and matching tasks, enhancing multi-modal alignment without object annotations. Li et al. [22] implemented a multi-modal mixture of encoder-decoder, **Bootstrapping Language-Image Pre-Training (BLIP)** network, which utilises noisy data from the web by bootstrapping the descriptions of images, where a captioner generates synthetic descriptions and a filter removes the noisy ones. While the above networks employ transformer architectures to apply cross-attention between image regions and descriptive words, they do not fully explore the relational reasoning within images.

2.4 Relational Reasoning Methods

Graphs are invaluable for representing and analysing relations [17, 18, 35]. In recent years, graph-based methods have shown an efficient way of reasoning with relations [3, 39, 40, 45, 50]. For a scene graph generation task, Lin et al. [27] explored the atom correlation-based graph propagation which incorporates prior knowledge in a more stable and comprehensive way, and Cuiet et al. [8] proposed a framework for visual relationship detection that uses word semantic and visual scene graphs to capture global context interdependency among object instances. For cross-modal information retrieval, Cao et al. [1] introduced a graph-based relation-aware attention module to weigh image fragments based on the pairwise relations of the fragments, and Li et al. [25] applied a GCN to extract relations between image regions and used the extracted relations to enhance image features. Chen et al. [4] proposed a **Two-Stream Hierarchical Similarity Reasoning (TSHSR)** network for image-text matching, utilising hierarchical similarity reasoning and a two-stream architecture to enhance matching by exploiting multi-level hierarchical similarity information and decomposing the matching into image-to-text and text-to-image levels.

3 Proposed VITR Network

Overview. The proposed VITR network is illustrated in Figure 2. Given an image I and a description D, VITR aims to embed the pair (I, D) into the shared latent space for predicting its similarity score $s(I, D)$. VITR comprises: (1) A text encoder which encodes the description D to incorporate pre-trained language knowledge. (2) A ViT encoder and a CNN-based local encoder which encode the image I and its regions as a global representation and a set of features respectively. (3) A relational reasoning module that represents image regions in relations and generates local representations of the image regions based on their descriptions. (4) A fusion module that predicts the similarity score $s(I, D)$ based on fusing the results of VITR's relational reasoning module and pre-trained knowledge using a sequence-optimised graph network.

3.1 Encoding the Description

VITR utilises a pre-trained text encoder (e.g., CLIP's text encoder [36] or pre-trained BERT [2]). This module encodes the description D as a global representation vector $u^{\text{glob}} \in \mathbb{R}^{d_1}$ and a collection of word embedding vectors $U = \{u_1, \ldots, u_n\}$, where n is the number of words in the description and $u_j \in \mathbb{R}^{d_1}$ is the jth word embedding vector with dimension d_1. The output of this module is (U, u^{glob}).

3.2 Encoding the Image

VITR encodes the image using two components. The ViT encoder utilises a pre-trained ViT network based on cross-modal learning, such as the image encoder of CLIP's ViT model. This module encodes an image I as a global representation vector $v^{\text{glob}} \in \mathbb{R}^{d_1}$. The local encoder utilises a pre-trained CNN to encode the image I into a set of regional representations:

$$\text{CNN}(I) = V = \{v_1, \ldots, v_k\},$$

where each feature $v_i \in \mathbb{R}^{d_2}$ encodes a salient region of the image and k is the total number of regions. The output of this module is (V, v^{glob}). Examples of such CNN networks include the image encoder of CLIP's ResNet model [36] or the ResNet backbone of Faster-RCNN [37].

3.3 Proposed Relational Reasoning

This section presents the methodology of the relational reasoning module. The comprehensive relational reasoning process within the module is visually represented in Figure 3. Firstly, the relational reasoning module aims to generate a relational matrix based on the pairwise relationships of image regions through the graph neural network. Then, the relational matrix subsequently weighs the features of the image regions, aiming to enhance the regions in relation. Finally, the alignment of the image regions with descriptive words ensures that only those image regions and their relationships associated with the descriptions are enhanced. The details and equations of the relational reasoning module are described as follows.

For further computation with crossing modalities, the elements of V and U are projected into a unified dimension d_3 as follows:

$$v_i^* = W^v(v_i),$$
$$u_j^* = W^u(u_j),$$

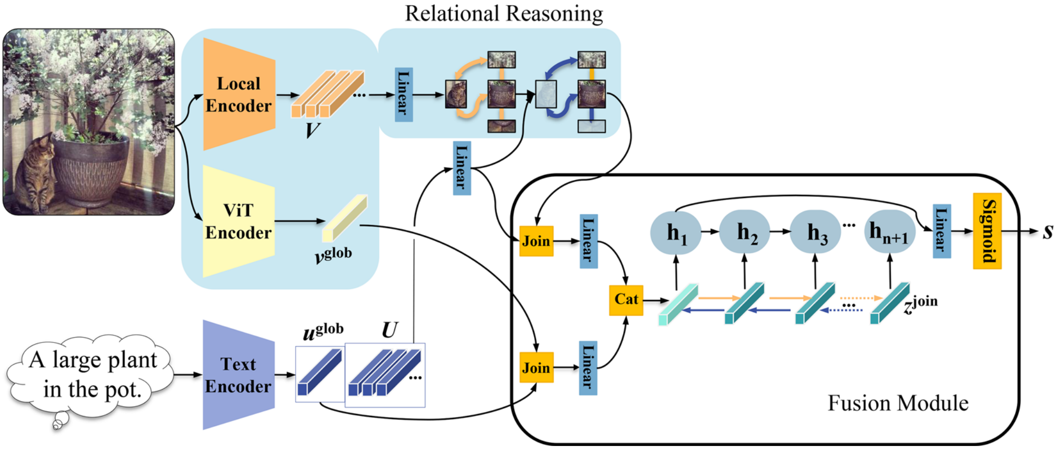


Fig. 2. An overview of the proposed VITR. VITR consists of: (1) a pre-trained *text encoder* that provide pre-trained language knowledge of an image's description; (2) a pre-trained *ViT encoder* that encodes an image as a global representation, and a CNN-based *local encoder* that extracts features from image regions; (3) a *relational reasoning* module that models the relations between regions in an image and generates local representations of the regions based on their descriptions; and (4) a *fusion* module that fuses the outputs from relational reasoning and pre-trained knowledge through a sequence-optimised graph network, and outputs the similarity score between the image I and description D.

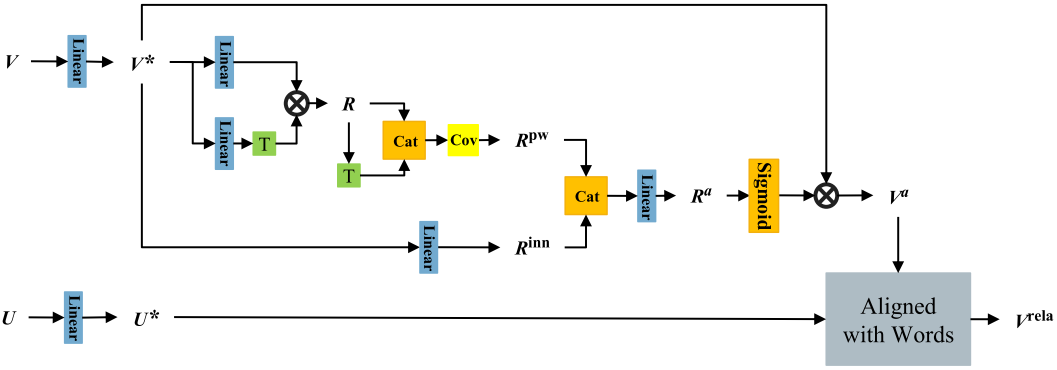


Fig. 3. Diagram of the relational reasoning module, where 'Cat', 'Cov', and 'T' denote the operations of concatenation, convolution, and transposition, respectively.

where the weight parameters W^v and W^u are both the fully connected layers with d_3 output neurons. Here, $v_i^* \in \mathbb{R}^{d_3}$ is the projected vector corresponding to the ith region. Similarly, $u_j^* \in \mathbb{R}^{d_3}$ is the jth embedding vector for the jth word in the description. Finally, set $V^* = \{v_1^*, \ldots, v_k^*\}$ and $U^* = \{u_1^*, \ldots, u_n^*\}$.

The regions of an image and their relationships are represented using a multi-layer graph neural network. Let $R \in \mathbb{R}^{k \times k}$ be a matrix of relations of regions and computed whose element (i, l), for any $1 \leq i, l \leq k$, is

$$[R]_{i,l} = W^{\varphi_1}(v_i^*)^{\mathrm{T}} W^{\phi_1}(v_l^*), \tag{1}$$

where the weight parameters W^{φ_1} and W^{ϕ_1} are both fully connected layers with d_3 output neurons. From this, a matrix of pairwise relations of regions is computed as

$$\underset{k\times k}{R^{\text{pw}}} = \sigma(W^{\text{R}_1}(\text{cat}[R^{\text{T}}, R])),$$

where cat denotes row-wise concatenation and σ denotes the tanh activation function. The weight parameter W^{R_1} is a One-Dimensional convolutional layer (kernel size 1; k output channels). Additionally, let $R^{\text{inn}} = (r_1^{\text{inn}}, \ldots, r_k^{\text{inn}}) \in \mathbb{R}^k$ hold the inner information for each vector in V^*:

$$r_i^{\text{inn}} = \sigma(W^{\text{R}_2}(v_i^*)) \in \mathbb{R}^k,$$

where the weight parameter W^{R_2} is a fully connected layer with one output neuron. Merge R^{pw} and R^{inn} into R^{a} as

$$\underset{k\times 1}{R^{\text{a}}} = \sigma(W^{\text{R}_3}(\text{cat}[R^{\text{pw}}, R^{\text{inn}}])),$$

where cat now denotes column-wise concatenation; the weight parameter W^{R_3} is a fully connected layer with one output neuron. A collection of representation vectors $V^{\text{a}} = \{v_1^{\text{a}}, \ldots, v_k^{\text{a}}\}$ for the k regions is then obtained as

$$v_i^{\text{a}} = \text{sigmoid}([R^{\text{a}}]_i)v_i^* \in \mathbb{R}^{d_3}, \tag{2}$$

where $[R^{\text{a}}]_i$ is ith element of R^{a}. Finally, Equations (1) and (2) can be recursively repeated $g_1 \in \mathbb{N}^+$ times ($g_1 = 4$ in this article). In this case, the output V^{a} from the repetition forms the input V^*.

Since not all visual vectors in the set V^{a} are relevant to the description, the visual vectors are weighted to generate local representations of the image that are aligned with the descriptive words, denoted $V^{\text{rela}} = \{v_1^{\text{rela}}, \ldots, v_n^{\text{rela}}\}$. Here, the newly generated image local representation vector aligned with the jth word is given by

$$v_j^{\text{rela}} = \sum_{i=1}^{k} a_{i,j} v_i^{\text{a}},$$

where

$$a_{i,j} = \frac{\exp(\gamma \bar{s}_{i,j})}{\sum_{i'=1}^{k} \exp(\gamma \bar{s}_{i',j})}, \tag{3}$$

are weights that are specified through a softmax function with inverse temperature parameter $\gamma > 0$ (set to $\gamma = 12$ by this article). Here,

$$\bar{s}_{i,j} = \frac{[\text{s}_{\text{cs}}(v_i^{\text{a}}, u_j^*)]_+}{\sqrt{\sum_{j'=1}^{n} [\text{s}_{\text{cs}}(v_i^{\text{a}}, u_{j'}^*)]_+^2}},$$

with $[x]_+ = \max(x, 0)$ is a normalised and thresholded version of the cosine similarity s_{cs}. In summary, the output of the relational reasoning module is (V^{rela}, U^*).

3.4 Proposed Fusion Module

This module predicts the similarity score $s(I, D)$ for the image–description pair by fusing the results of the relational reasoning module and the global representations of the image and description. The process is described as follows.

The local and global image-description representations are combined and embedded in the same low-dimensional latent space (intended to reduce computational complexity) for fusion processing.

More formally, a vector z^{glob} for joining a global image–description representation pair $(v^{\text{glob}}, u^{\text{glob}})$ and vectors z_j^{rela} for joining local image–description representation pairs $(v_j^{\text{rela}}, u_j^*)$ are computed as

$$z^{\text{glob}} = (u^{\text{glob}} - v^{\text{glob}})^2,$$
$$z_j^{\text{rela}} = (u_j^* - v_j^{\text{rela}})^2,$$

where $(\,\cdot\,)^2$ is applied element-wise. Furthermore, define the vectors (in $\mathbb{R}^{d_4}$):

$$z_0^{\text{join}} = W^{\text{glob}}(z^{\text{glob}}),$$
$$z_i^{\text{join}} = W^{\text{rela}}(z_i^{\text{rela}}), \quad \text{for} i = 1, \ldots, n,$$

where the weight parameters W^{rela} and W^{glob} are both fully connected layers with d_4 (e.g., 128) output neurons.

To ensure that z_j^{join} contains sufficient contextual information, it can be treated as a node for constructing a graph. The edge matrix $E = \mathbb{R}^{(n+1)\times(n+1)}$ is obtained (for any $1 \leq j, l \leq (n+1)$) as

$$[E]_{j,l} = W^{\varphi_2}(z_j^{\text{join}})^{\mathrm{T}} W^{\phi_2}(z_l^{\text{join}}), \tag{4}$$

where the weight parameters W^{φ_2} and W^{ϕ_2} are both fully connected layers with d_4 output neurons. Then the information among the joined vectors is fused as

$$z_j^{\text{fuse}} = \sum_{l=1}^{n+1} W^{\text{fuse}}(\text{sigmoid}([E]_{j,l}) z_j^{\text{join}}), \tag{5}$$

where the weight parameter W^{fuse} is a fully connected layer with d_4 output neurons. The jth fused representation, z_j^{fuse}, is computed by aggregating weighted sums of the jth joined representation, z_j^{join}, and the sigmoid-activated elements of the edge matrix E. Finally, set $Z^{\text{fuse}} = \{z_1^{\text{fuse}}, \ldots, z_{n+1}^{\text{fuse}}\}$. Equations (4) and (5) can be recursively repeated $g_2 \in \mathbb{N}^+$ times ($g_2 = 2$ in this article), where the output Z^{fuse} from the last time is taken as the input for the next time. A sequence optimiser is utilised to dynamically capture and incorporate the temporal dependencies among the elements of Z^{fuse}. This allows the module to generate a rich and complex combined representation of Z^{fuse} as

$$\{h_j\}_{j=1}^{n+1} = \text{GRU}(Z^{\text{fuse}}),$$

where $\{h_j\}_{j=1}^{n+1}$ are the hidden states of a GRU layer, and only h_1 is taken as the combined representation of Z^{fuse}.

Finally, the similarity score s for a pair (I, D) is predicted as

$$s(I, D) = \text{sigmoid}(W^{\text{h}}(h_1)),$$

where the weight parameter W^{h} is a fully connected layer with one output neuron.

3.5 Training VITR

The pre-trained models—text, ViT, and local encoders—constitute an integral part of the VITR framework. The remaining parameters within VITR undergo a collaborative training process facilitated by LSEH [14]. LSEH, which serves as an advanced version of the hard negatives loss function, focuses on learning the distances between image–description pairs [11]. Consider $\{(I_1, D_1), \ldots, (I_m, D_m)\}$ as a training dataset consisting of image–description pairs. Each image I_p is associated with its corresponding relevant description D_p, where p denotes the pair index and m represents the total number of pairs in the training set. Given a relevant image–description pair (I_p, D_p), the result of LSEH only takes the max from the irrelevant pairs as

$$\begin{aligned}\mathrm{L}(I_p, D_p) = &\max_{\hat{D}_p}[\alpha + \lambda\, \mathrm{s_{cs}}(D_p^{'}, \hat{D}_p^{'}) + s(I_p, \hat{D}_p) - s(I_p, D_p)]_+ \\ &+ \max_{\hat{I}_p}[\alpha + \lambda\, \mathrm{s_{cs}}(D_p^{'}, \hat{D}_p^{'}) + s(D_p, \hat{I}_p) - s(I_p, D_p)]_+,\end{aligned}$$

where α is a margin parameter, λ is a temperature parameter, and $\hat{D}_p$ and $\hat{I}_p$ are irrelevant images and descriptions, respectively (from a mini-batch). The semantic factors $\lambda\, \mathrm{s_{cs}}(D_p^{'}, \hat{D}_p^{'})$ dynamically adjust the margin α according to the cosine similarity between $D_p^{'}$ and $\hat{D}_p^{'}$ for flexible learning of the network.

Let $D_p^{'}$ and $\hat{D}_p^{'}$ represent the decomposition eigenvalues derived from D_p and $\hat{D}_p$, respectively, and $D_p^{'}$ and $\hat{D}_p^{'}$ are obtained through the application of truncated **Singular Value Decomposition (SVD)** to the constructed description matrix A. The specific operations are as follows: Define the matrix $A = \mathrm{cat}[D_1^{\mathrm{T}}, \ldots, D_m^{\mathrm{T}}] \in \mathbb{R}^{m \times w}$, where m is the number of descriptions in the training dataset, w denotes the dimension of each description, and cat denotes row-wise concatenation. Then truncated SVD is applied to A as follows:

$$\underset{m\times w}{A} \approx \underset{m\times d_5}{X}\ \underset{d_5\times d_5}{\Lambda}\ \underset{d_5\times w}{Y^{\mathrm{T}}}, \quad \underset{m\times d_5}{B} = \underset{m\times w}{A}\ \underset{w\times d_5}{Y},$$

where d_5 is the number of singular values. The m rows of matrix B consist of the vectors $D_1^{'}, \ldots, D_m^{'} \in \mathbb{R}^{d_5}$, representing the decomposition eigenvalues derived from descriptions $D_1, \ldots, D_m$. Finally, $D_p^{'}$ and $\hat{D}_p^{'}$ are selected from the set $\{D_1^{'}, \ldots, D_m^{'}\}$. As recommended by [14], this article sets the margin parameter α to 0.185 and the temperature parameter λ to 0.025.

3.6 Proposed Turbo Module for Improving Retrieval Efficiency

Inspired by Li et al. [23], which presents a strategy for candidate selection to enhance the efficiency of cross-modal information retrieval, this article extends this idea by proposing a turbo module. Comprehensive experiments are planned to evaluate the turbo module's impact on both retrieval accuracy and computational speed, as detailed in Section 4.5. The global similarity was optimised during the pre-training stage; therefore, VITR does not need to optimise it further. Consequently, the turbo module (which is not for training) is based on using the global similarity. The turbo module is shown in Figure 4.

Input. The turbo receives the images, the output of the ViT encoder (i.e., the global representation of the images), and the output of the text encoder (i.e., the global representations of the descriptions).

Operation. The turbo ranks the descriptions based on the cosine similarities between the query image's global representation and the global representations of the descriptions for image-to-text retrieval or ranks the images based on the cosine similarities between the query description's global representation with the images' global representations for text-to-image retrieval. It then selects the top N ($N \in \mathbb{N}^+$) relevant candidates for the query based on the ranking results.

Output. For image-to-text retrieval, turbo sends the candidate descriptions' word embeddings and global representations to the relational reasoning and fusion modules, respectively, and the query image and its global representation to the local encoder and the fusion modules, respectively. For text-to-image retrieval, turbo sends the candidate images and their global representations to the local encoder and the fusion module, respectively and the query description's word embeddings and global representation to the relational reasoning and fusion modules, respectively.

Finally, each module in VITR performs computations based on the received results from turbo to finalise the ranking of candidate descriptions or images for the query. By using the turbo module,

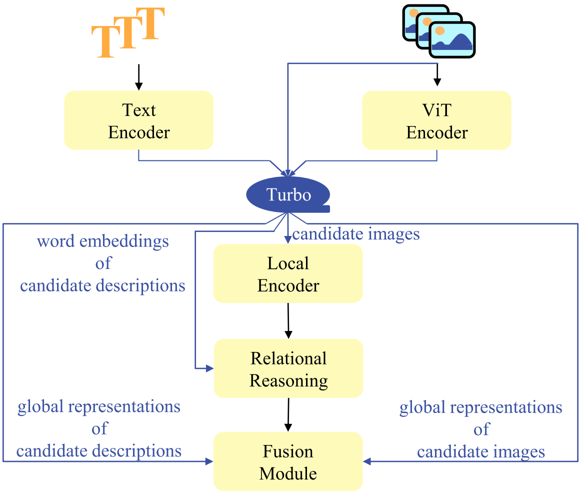

TTT
Text Encoder
ViT Encoder
Turbo
word embeddings of candidate descriptions
candidate images
Local Encoder
Relational Reasoning
global representations of candidate descriptions
global representations of candidate images
Fusion Module

Fig. 4. The proposed *turbo* module for VITR selects the top N candidate descriptions or images for the query and sends the necessary candidates' embeddings or representations to the relevant modules for further finalisation of ranking.

the computational complexity of the major components of VITR (involving the local encoder, and the relational reasoning and fusion modules) is reduced by a factor of $\frac{length}{N}$, where $length$ is the length of the database and N is the number of candidates.

4 Experiments

The proposed VITR underwent evaluation using the Flickr30K [46], MS-COCO [28], RefCOCOg [33], and CLEVR [16] datasets for image-to-text and text-to-image retrieval tasks. The performance of VITR was then benchmarked against that of state-of-the-art networks.

4.1 Evaluation Measures and Datasets

The evaluation metric used for the cross-modal information retrieval experiments is Recall at rank K (Recall@K), which measures the percentage of relevant items included in the top K retrieved results [15]. The experiments aim to evaluate the network's ability to retrieve at least one relevant item from a given list of relevant items, and the average Recall is computed across the results of the evaluated queries. The Flickr30K, MS-COCO, RefCOCOg, and CLEVR datasets are split as shown in Table 1 and described as follows.

The *Flickr30K* [46] and *MS-COCO* [28] datasets are the commonly used benchmark for evaluating the performance of VSE networks [2, 25, 36]. Each image in the datasets of Flickr30K and MS-COCO is associated with five textual descriptions.

The *RefCOCOg* dataset [33] contains real-world images sourced from the MS-COCO dataset [28], and their corresponding descriptions, provided by the University of Maryland, focus on the relations between regions within the images. On average, each image in the dataset is associated with four relevant descriptions. RefCOCOg is commonly used for referring expression tasks, and it is also suitable for image-to-text and text-to-image retrieval tasks, especially when evaluating networks that focus on the relations expressed in the queries. For image-to-text retrieval, an image from RefCOCOg is used as the query to retrieve relevant textual descriptions. Conversely, for text-to-image retrieval, a textual description from RefCOCOg serves as the query to retrieve relevant images.

Table 1. Dataset Split of Flickr30K, MS-COCO, RefCOCOg, and CLEVR

Dataset	Modality	Train	Validate	Test
Flickr30K	Images	29,000	1,000	1,000
	Descriptions	145,000	5,000	5,000
MS-COCO	Images	113,287	1,000	5,000
	Descriptions	566,435	5,000	25,000
RefCOCOg	Images	21,899	1,300	2,600
	Descriptions	80,512	4,896	9,582
CLEVR	Images	30,000	1,000	1,000
	Descriptions	98,345	3,136	3,121

The *CLEVR* dataset [16] consists of images depicting Three-Dimensional-rendered objects. Since this dataset has not been specifically tailored for relation-focused cross-modal retrieval tasks, image descriptions were formulated using the given relational annotations like 'left', 'right', 'front', and 'behind'. The dataset was then split into train, test, and validation sets. On average, each CLEVR image is associated with three relevant descriptions. An example description is 'A large blue metal cube is behind a large blue rubber sphere'.

4.2 Implementation Details

All experiments were conducted on a workstation with NVIDIA RTX3090 GPU with PyTorch framework, and the source code files are provided in our GitHub repository.[1] The networks were implemented as follows. Four CLIP baseline models were selected. These were the base ViT model ('ViT-B/16' with dimension d_1 of 512), the large ViT models ('ViT-L/14' and 'ViT-L/14@336px' with dimension d_1 of 768), where 'ViT-L/14@336px' is 'ViT-L/14' pre-trained at a high 336-pixel resolution to enhance the retrieval performance, and the ResNet101 model ('RN101') denoted as CLIP_{B16}, CLIP_{L14}, $\text{CLIP}_{\text{L14px}}$, and $\text{CLIP}_{\text{RN101}}$, respectively [36]. Each model was fine-tuned for each dataset to present its best performance, and the hyperparameter settings follow each model's benchmark settings [36]. VITR was implemented using the ViT and text encoders from the fine-tuned CLIP_{B16}, CLIP_{L14}, and $\text{CLIP}_{\text{L14px}}$ models, this resulted in three models of VITR network, namely VITR_{B}, VITR_{L}, and VITR_{Lpx}, respectively. The image encoder of the fine-tuned $\text{CLIP}_{\text{RN101}}$ was applied for encoding image regions for VITR_{B}, VITR_{L}, and VITR_{Lpx}, and it extracts 49 features (with dimension d_2 of 2,048) of regions from each image. VITR_{B}, VITR_{L}, and VITR_{Lpx} underwent training on each dataset for 20 epochs, with a set batch size of 128. The learning rate was fixed at 0.0004 and was subjected to a decay rate of 0.1, commencing at the 5th epoch. This training process made use of the Adam optimiser.

4.3 Comparison of Cross-Modal Information Retrieval Performance on Benchmarks

This section focuses on the comparison of VITR with various state-of-the-art networks on the benchmark datasets Flickr30K and MS-COCO. The results on the Flickr30K and MS-COCO datasets are shown in Table 2. VITR_{Lpx} achieved a Recall@1 of 96.4% for image-to-text and 86.3% for text-to-image retrieval on Flickr30K. On MS-COCO, VITR_{Lpx}'s Recall@1 was 77.9% for image-to-text and 60.3% for text-to-image retrieval. The comparison with other networks is summarised as follows.

(1) Compared to ALBEF, VITR_{Lpx}'s Recall@1 was higher by 0.5% for image-to-text and by 0.7% for text-to-image retrieval on Flickr30K, and by 0.3% for image-to-text retrieval on MS-COCO. The

[1] https://github.com/yangong23/VITR

Table 2. Average Recall@K (R@K) Values (%) of Cross-Modal Information Retrieval Networks on the Test Sets of Flickr30K and MS-COCO

Network	#Parameters	Flickr30K						MS-COCO					
		Image-to-Text			Text-to-Image			Image-to-Text			Text-to-Image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
VSE++ [11]	-	52.9	80.5	87.2	39.6	70.1	79.5	41.3	71.1	81.2	30.3	59.4	72.4
PFAN++ [42]	-	70.1	91.8	96.1	52.7	79.9	87.0	51.2	84.3	89.2	41.4	70.9	79.0
TSHSR [4]	-	76.3	93.0	95.8	56.6	81.2	85.9	57.4	-	91.4	41.4	-	81.2
SGRAF [10]	-	77.8	94.1	97.4	58.5	83.0	88.8	57.8	91.6	41.9	81.3		
VSRN++ [25]	-	79.2	94.6	97.5	60.6	85.6	91.4	54.7	82.9	90.9	42.0	72.2	82.7
Unicoder [20]	-	86.2	96.3	99.0	71.5	90.9	94.9	62.3	87.1	92.8	46.7	76.0	85.3
UNITER [5]	303M	87.3	98.0	99.2	75.6	94.1	96.8	63.3	87.0	93.1	48.4	76.7	85.9
ERNIE-ViL [47]	88.7M	98.0	99.2	76.7	93.6	96.4	-	-	-	-	-	-	
ViSTA-L [6]	-	89.5	98.4	99.6	75.8	94.2	96.9	63.9	87.1	93.0	47.4	75.0	84.0
VILLA [13]	-	87.9	97.5	98.8	76.3	94.2	96.8	-	-	-	-	-	-
CLIP$_{\mathrm{RN101}}$	120M	88.3	98.2	99.4	72.9	92.6	96.2	66.9	87.8	93.6	48.6	75.4	84.4
VSE∞ [2]	-	88.7	98.9	99.8	76.1	94.5	97.1	68.1	90.2	95.2	52.7	80.2	88.3
CLIP$_{\mathrm{B16}}$	150M	91.2	98.9	99.4	77.0	94.1	97.4	70.0	90.7	95.4	53.0	80.5	88.7
COTS† [30]	-	91.7	99.0	99.9	78.3	94.9	97.2	70.6	91.0	95.3	53.7	80.2	87.8
CLIP$_{\mathrm{L14}}$	428M	92.6	99.2	99.6	77.8	95.2	97.7	74.6	92.3	95.8	57.3	81.6	88.9
CLIP$_{\mathrm{L14px}}$	428M	94.9	99.6	99.8	83.9	97.4	98.8	75.8	92.7	96.3	58.3	82.4	89.4
ALBEF [23]	223M	95.9	99.8	100.0	85.6	97.5	98.9	77.6	94.3	97.2	60.7	84.3	90.5
MVPTR [26]	-	95.2	99.7	100.0	84.0	96.8	98.5	77.3	93.6	96.9	60.1	84.0	90.7
RVLM [48]	-	95.6	99.8	100.0	85.7	97.6	98.8	79.5	95.1	97.9	63.1	85.6	91.9
BLIP [22]	446M	97.4	99.8	99.9	87.6	97.7	99.0	82.4	95.4	97.9	65.1	86.3	91.8
VITR$_{\mathrm{B}}$	246M	93.7	99.1	99.8	80.8	95.7	97.9	72.7	91.0	95.4	55.2	80.2	88.0
VITR$_{\mathrm{L}}$	524M	94.7	99.7	99.9	82.5	96.7	98.3	76.4	93.2	96.4	58.8	82.6	89.6
VITR$_{\mathrm{Lpx}}$	524M	96.4	99.8	100.0	86.3	97.7	99.1	77.9	93.7	96.9	60.3	83.6	90.2

COTS† denotes the ensemble results of two models.
The @ symbol in the table represents the Recall at a specific value of K (R@K).

metric where VITR$_{\mathrm{Lpx}}$ performed lower than ALBEF was in text-to-image retrieval on MS-COCO, with a decrease of 0.4%.

(2) VITR$_{\mathrm{Lpx}}$ outperformed MVPTR's Recall@1 by 1.2% and 2.3% for image-to-text and text-to-image retrieval, respectively, on Flickr30K, and by 0.6% and 0.2% for image-to-text and text-to-image retrieval, respectively, on MS-COCO.

(3) In comparison to RVLM, VITR$_{\mathrm{Lpx}}$'s Recall@1 was higher by 0.8% for image-to-text and by 0.6% for text-to-image retrieval on Flickr30K. On MS-COCO, VITR$_{\mathrm{Lpx}}$'s Recall@1 was lower than RVLM's by 1.6% for image-to-text and by 2.8% for text-to-image retrieval, respectively.

(4) VITR$_{\mathrm{Lpx}}$'s Recall@1 is close to BLIP's, being lower by only 1.0% for image-to-text and 1.3% for text-to-image retrieval on Flickr30K, and by 4.5% and 4.8% for image-to-text and text-to-image retrieval, respectively, on MS-COCO. Despite VITR$_{\mathrm{Lpx}}$'s lower performance in cross-modal information retrieval compared to BLIP, it achieved faster retrieval time. In tests of image-to-text and text-to-image retrieval across 1,000 images and 5,000 descriptions (using the Flickr30K test set), BLIP required 1,430.4 s to complete the tasks. In contrast, VITR$_{\mathrm{Lpx}}$, without the turbo module and hence no re-ranking strategy, needed only 116.7 s.

4.4 Further Analysis of VITR on Extended Datasets

Section 4.3 compared VITR with state-of-the-art models on the commonly used benchmark datasets Flickr30K and MS-COCO. This section provides a further analysis of VITR's performance on extended datasets, specifically the relation-focused RefCOCOg and CLEVR datasets. VSRN++, VSE∞, and CLIP were selected as the baseline models. The results are shown in Table 3.

Table 3. Average Recall@K Values (%) of Cross-Modal Information Retrieval Networks on the Test Sets of RefCOCOg and CLEVR

	RefCOCOg						CLEVR					
Network	Image-to-Text			Text-to-Image			Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
VSRN++	20.0	44.9	57.3	13.8	34.6	47.8	64.3	96.9	99.4	58.5	91.0	96.1
VSE∞	31.1	58.3	69.7	19.5	42.8	55.2	67.8	99.7	99.9	70.8	99.0	99.5
CLIP$_{\mathrm{RN101}}$	36.3	61.3	71.2	20.8	44.2	56.7	65.4	98.1	99.8	61.8	95.7	98.1
CLIP$_{\mathrm{B16}}$	39.3	64.3	75.0	23.8	48.4	60.4	66.8	98.6	100.0	65.5	97.8	99.5
CLIP$_{\mathrm{L14}}$	42.4	65.5	75.1	25.2	48.9	60.4	65.6	99.4	99.9	65.2	97.6	99.1
CLIP$_{\mathrm{L14px}}$	44.0	67.8	76.9	27.4	51.7	63.4	69.6	99.6	99.9	67.8	98.6	99.4
VITR$_{\mathrm{B}}$	42.9	68.2	79.2	27.9	53.5	65.6	88.3	99.7	100.0	79.4	99.4	99.8
VITR$_{\mathrm{L}}$	45.2	71.1	80.5	29.5	55.1	66.8	90.7	99.9	99.9	79.3	99.5	99.8
VITR$_{\mathrm{Lpx}}$	48.7	72.9	81.6	30.6	55.7	67.1	93.8	99.9	100.0	86.3	99.5	99.8

(1) Results on RefCOCOg. VITR$_{\mathrm{Lpx}}$ reached a Recall@1 of 48.7% for image-to-text and a Recall@1 of 30.6% for text-to-image retrieval. Observing the performance of the networks using the Recall@1 metric, VITR$_{\mathrm{Lpx}}$ outperformed CLIP$_{\mathrm{L14px}}$ by 4.7% and 3.2% for image-to-text and text-to-image retrieval, respectively and also outperformed VSE∞ by 17.6% and 11.1% for those tasks, respectively. VITR$_{\mathrm{B}}$ reached a Recall@1 of 42.9% and 27.9% for image-to-text and text-to-image retrieval, respectively and outperformed CLIP$_{\mathrm{B16}}$ by 3.6% and 4.1% for those tasks, respectively.

(2) Results on CLEVR. For VITR$_{\mathrm{Lpx}}$, Recall@1 reached 93.8% for image-to-text and 86.3% for text-to-image retrieval and outperformed CLIP$_{\mathrm{L14px}}$ by 24.2% and 18.5% for those tasks, respectively. VITR$_{\mathrm{Lpx}}$'s Recall@1 also outperformed VSE∞'s Recall@1 by 26.0% and 15.5% for image-to-text and text-to-image retrieval, respectively. The Recall@1 values of VITR$_{\mathrm{B}}$ were 88.3% and 79.4% for image-to-text and text-to-image retrieval, respectively. The Recall@1 values of VITR$_{\mathrm{B}}$ outperformed that of CLIP$_{\mathrm{B16}}$ by 21.5% for image-to-text and 13.9% for text-to-image retrieval, respectively.

4.5 Results of VITR Using the Turbo Module

Comparison of Retrieval Time between VITR with and without Turbo. Table 4 compares the retrieval times of VITR$_{\mathrm{L}}$ with and without the turbo module, when these are applied to the RefCOCOg test set. Here, N is the number of selected candidates by turbo (see Section 3.6), in the table. For retrieval of relevant descriptions from a pool of 9,582 using a single query image, the average retrieval time of VITR$_{\mathrm{L}}$ with turbo ($N = 200$) is 0.3 s, which is 13.7 s faster than that without turbo and 10.5 s faster than that of UNITER. For retrieval of relevant images from a pool of 2,600 using a single query description, the average retrieval time of VITR$_{\mathrm{L}}$ with turbo ($N = 200$) is 0.1 s, which is 1.7 s faster than that without turbo and 4.6 s faster than that of UNITER. Compared to the one-stage VSE network CLIP$_{\mathrm{L14}}$, VITR$_{\mathrm{L}}$ without turbo is 13.05 s and 1.76 s slower in image-to-text and text-to-image retrieval, respectively. However, VITR$_{\mathrm{L}}$ with turbo ($N = 200$) is only 0.25 s and 0.06 s slower than CLIP$_{\mathrm{L14}}$ in image-to-text and text-to-image retrieval, respectively. VITR$_{\mathrm{L}}$ with turbo outperforms CLIP$_{\mathrm{L14}}$ in cross-modal information retrieval performance with acceptable time efficiency.

The Retrieval Performance of VITR with Turbo. This section evaluates the impact of the proposed turbo module on the retrieval performance of VITR using the RefCOCOg test set. Table 5 shows that, for image-to-text and text-to-image retrieval, the retrieval performance of VITR using turbo with N set to 200 and 500 is the same as that of VITR without turbo. When N is set to 100, VITR with turbo underperformed VITR without turbo with a difference of 0.2% for image-to-text retrieval

Table 4. Comparison of the Retrieval Time of Different Models, Including VITR$_L$ with and without Turbo, and UNITER, Using the RefCOCOg Test Set

Task	CLIP$_{L14}$	VITR$_L$ (Turbo $N = 200$)	VITR$_L$ (Turbo $N = 500$)	VITR$_L$ (Without Turbo)	UNITER
Image-to-Text	0.05s	0.3 s	0.8 s	14.0 s	10.8 s
Text-to-Image	0.04s	0.1 s	0.3 s	1.8 s	4.7 s

Table 5. Results of VITR$_L$ with Turbo for Cross-Modal Information Retrieval on the RefCOCOg Test Set

Turbo N	Image-to-Text			Text-to-Image		
	Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10
100	45.2	71.1	80.3	29.5	55.1	66.7
200	45.2	71.1	80.5	29.5	55.1	66.8
500	45.2	71.1	80.5	29.5	55.1	66.8
Without Turbo	45.2	71.1	80.5	29.5	55.1	66.8

Table shows average Recall@K values (%).

Table 6. Results of Ablation Studies on VITR's Variant Networks for Cross-Modal Information Retrieval on the RefCOCOg Test Set

Network	Method	Image-to-Text			Text-to-Image		
		Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10
VSE∞	Baseline	31.1	58.3	69.7	19.5	42.8	55.2
CLIP$_{L14}$	Baseline	42.4	65.5	75.1	25.2	48.9	60.4
VITR$_\infty$	VSE∞'s encoders	36.9	63.8	74.7	25.1	50.3	61.8
VITR-NoViT	Remove ViT	36.1	61.6	72.2	24.3	49.1	60.8
VITR-NoRel	Remove RR	43.1	66.7	76.9	25.3	49.3	60.4
VITR	Original	45.2	71.1	80.5	29.5	55.1	66.8

Table shows average Recall@K values (%).

and 0.1% for text-to-image retrieval on Recall@10. The results in Table 4 and 5 suggest that VITR with the proposed turbo ($N \geqslant 200$) achieved the same retrieval performance as VITR without turbo, but in a faster retrieval time.

4.6 Ablation Studies on the Fusion

This section undertakes a series of ablation studies to assess the influence of integrating pre-trained knowledge and the results of relational reasoning within the proposed VITR network. Experiments were carried out by creating variants of VITR (VITR$_L$ model) and applying those to the RefCOCOg test set.

The first experiment aims to evaluate the performance of VITR when it does not utilise the ViT's image global representation, thereby assessing the impact of fusing the image global representation using the fusion module on the network. For this experiment, a new variant of VITR was created, namely VITR-NoViT, that removes the ViT encoder. As shown in Table 6, VITR-NoViT outperformed VSE∞ for image-to-text and text-to-image on Recall@1, with average improvements of 5.0% and

Table 7. Average Recall@K Values (%) of the Ablation Studies on the Number of Graph Layers (g_2) in the Fusion Module of VITR for Cross-Modal Information Retrieval on the RefCOCOg Test Set

Number of Layers	Image-to-Text			Text-to-Image		
	Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10
$g_2 = 0$	44.1	70.5	80.0	29.3	55.0	66.7
$g_2 = 1$	44.7	70.8	80.4	29.3	54.9	66.8
$g_2 = 2$	45.2	71.1	80.5	29.5	55.1	66.8

4.8%, respectively. In addition, VITR-NoViT underperformed VITR by 9.1% for image-to-text retrieval and 5.2% for text-to-image retrieval on Recall@1.

The second experiment aims to evaluate the performance of VITR when the relational reasoning module is removed, thereby assessing the impact of excluding the results of relational reasoning fusion on the network. For this experiment, a new variant of VITR was created, namely VITR-NoRel that does not include the relational reasoning module. The relational reasoning module was replaced by two GRUs, one for pooling the text and another for pooling the region features of images. As shown in Table 6, the results of Recall@1 of VITR-NoRel outperformed that of CLIP$_{L14}$ by 0.7% and 0.1%, respectively, and the results suggest that the observed improvement is a result of ViT's pre-trained global knowledge being integrated into the network along with the results obtained from GRUs. Furthermore, it was observed that the performance of VITR-NoRel was worse than that of VITR by 2.1% and 4.2% for Recall@1 in image-to-text and text-to-image retrieval tasks, respectively.

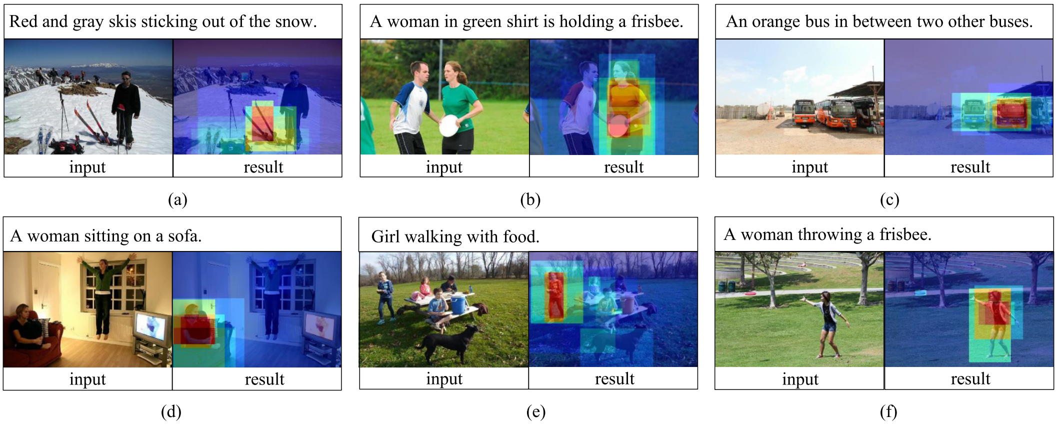
The third experiment aims to evaluate VITR's performance using the encoders from VSE∞, thereby assessing its adaptability with encoders from one-stage VSE networks beyond CLIP. For this experiment, a new variant of VITR was created, namely VITR$_\infty$, which replaces VITR's ViT encoder and text encoder with VSE∞'s image encoder and text encoder, respectively. As shown in Table 6, the Recall@1 results of VITR∞ outperformed VSE∞ by 5.8% and 5.6% for image-to-text and text-to-image retrieval, respectively. The results suggest that one-stage VSE networks such as VSE∞ can enhance their cross-modal information retrieval performance due to fusing VITR's relational reasoning results.

The fourth experiment aims to verify the effectiveness of the proposed VITR with simpler designs. The number of graph layers (g_2) in the fusion module of VITR varied from 0 to 2 while keeping the same input features (i.e., the global and local representations encoded by CLIP), as shown in Table 7. For image-to-text retrieval, VITR's Recall@1 increases from 44.1% when $g_2 = 0$ to 45.2% when $g_2 = 2$. For text-to-image retrieval, VITR's Recall@1 also shows a rise, starting at 29.3% for $g_2 = 0$ and increasing to 29.5% for $g_2 = 2$. The progression in Table 7 suggests a positive correlation between the number of graph layers (g_2) in the fusion module of VITR and the Recall@1 performance.

The last experiment aims to verify the effectiveness of the proposed VITR with variations in the number of image regions (k) used for processing. VITR was tested with setting k to 21, 28, 35, 42, and 49, respectively. The results, as shown in Table 8, indicate a progressive improvement in Recall@1, Recall@5, and Recall@10 metrics for both image-to-text and text-to-image retrieval tasks as the number of image regions increases. Starting from $k = 21$, with Recall@1 values of 42.6% for image-to-text and 26.8% for text-to-image, the performance steadily enhances, reaching peak values of 45.2% for image-to-text and 29.5% for text-to-image at $k = 49$. The results of Table 8 suggest that VITR benefits from the richer representation provided by higher numbers of regions, effectively leveraging the detailed spatial information to improve the alignment between textual descriptions and visual content.

Table 8. Average Recall@K Values (%) of the Ablation Studies on the Number of Image Regions (k) Used by VITR for Cross-Modal Information Retrieval on the RefCOCOg Test Set

Number of Image Regions	Image-to-Text			Text-to-Image		
	Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10
$k = 21$	42.6	68.2	77.7	26.8	51.2	63.4
$k = 28$	43.2	68.5	78.6	27.0	52.4	64.7
$k = 35$	44.4	70.3	79.5	28.3	53.7	65.5
$k = 42$	44.6	70.5	80.5	29.5	54.9	66.8
$k = 49$	45.2	71.1	80.5	29.5	55.1	66.8


Red and gray skis sticking out of the snow.
input
result
A woman in green shirt is holding a frisbee.
input
result
An orange bus in between two other buses.
input
result
A woman sitting on a sofa.
input
result
Girl walking with food.
input
result
A woman throwing a frisbee.
input
result

(a) (b) (c) (d) (e) (f)

Fig. 5. Visually representing the relational reasoning performance of VITR. In this figure, given a textual query and an input image, the visualisation is generated by highlighting the relevant image regions and darkening the irrelevant image regions.

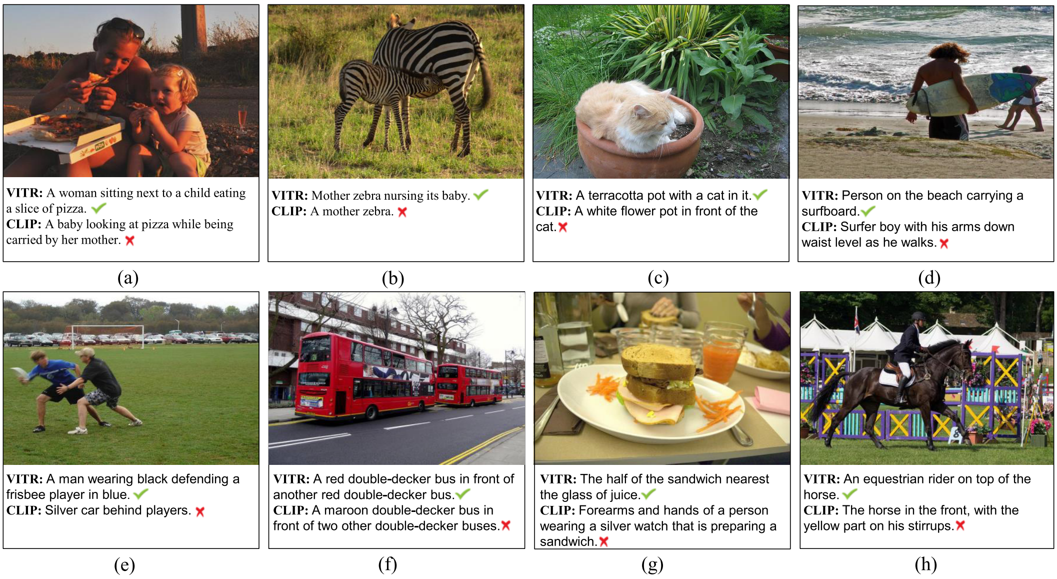
5 Visualisation

5.1 Visually Representing the Relational Reasoning Performance of VITR

CLIP does not explicitly model the relationships between image regions due to the lack of the relational reasoning component. Consequently, this section only centers on showcasing the performance of VITR's relational reasoning module, illustrating the capabilities of VITR in understanding and modeling image regions' relationships. Figure 5 presents an example visualisation of relational reasoning generated by the proposed VITR. In Figure 5, the heat map highlights the image regions relevant to the textual query, and it is generated by the relational reasoning module as follows.

Set $\{a_{i1}, \ldots, a_{in}\}$, see Equation (3), holds the weights for the ith image region, so let $\bar{a}_i$ denote the average value the set. Let set $\{\bar{a}_1, \ldots, \bar{a}_k\}$ holds the values of all image regions, and let its min-max normalisation result be $\{\bar{a}'_1, \ldots, \bar{a}'_k\} \in [0, 1]$. Therefore, $\bar{a}'_i$ is used as the heat degree for the ith image region.

The images from Figure 5(a)–(f) show that the image regions only received focus by the relational reasoning module when they were mentioned in the query description. For example, in Figure 5(b), the image regions relevant to ⟨'woman', 'holding', 'frisbee'⟩ were the focus, while the other main region 'man' in the image was ignored because it is irrelevant to the query description. The results of Figure 5 visually show the relational reasoning performance in VITR.


VITR: A woman sitting next to a child eating a slice of pizza. ✓
CLIP: A baby looking at pizza while being carried by her mother. ✗

(a)

VITR: Mother zebra nursing its baby. ✓
CLIP: A mother zebra. ✗

(b)

VITR: A terracotta pot with a cat in it. ✓
CLIP: A white flower pot in front of the cat. ✗

(c)

VITR: Person on the beach carrying a surfboard. ✓
CLIP: Surfer boy with his arms down waist level as he walks. ✗

(d)

VITR: A man wearing black defending a frisbee player in blue. ✓
CLIP: Silver car behind players. ✗

(e)

VITR: A red double-decker bus in front of another red double-decker bus. ✓
CLIP: A maroon double-decker bus in front of two other double-decker buses. ✗

(f)

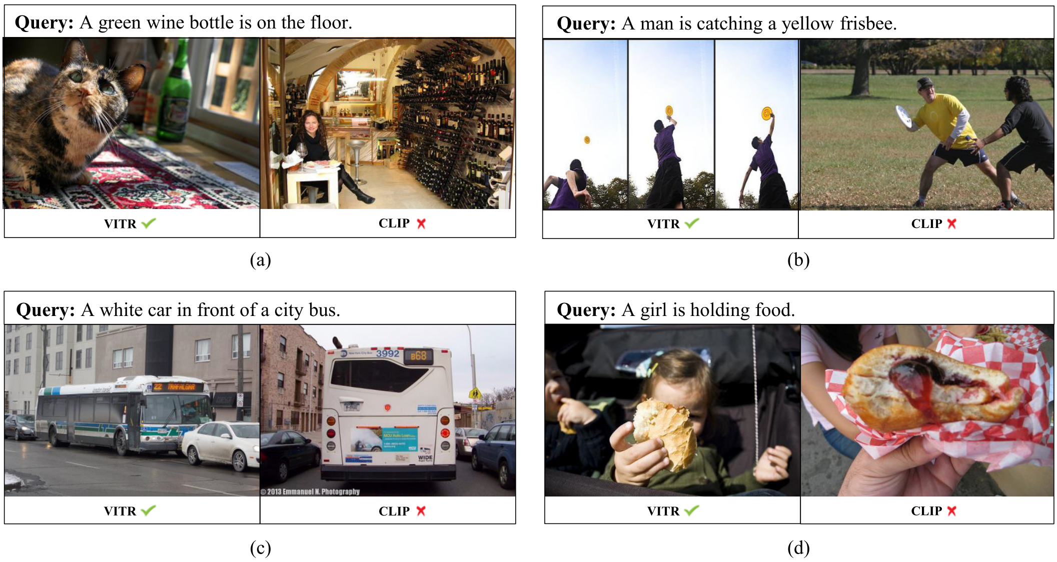
VITR: The half of the sandwich nearest the glass of juice. ✓
CLIP: Forearms and hands of a person wearing a silver watch that is preparing a sandwich. ✗

(g)

VITR: An equestrian rider on top of the horse. ✓
CLIP: The horse in the front, with the yellow part on his stirrups. ✗

(h)

Fig. 6. A comparison of the top one results for image-to-text retrieval using CLIP and VITR. CLIP's retrieved descriptions including the details do not match the query image, while VITR's retrieved descriptions concentrate on specific details of the image.


Query: A green wine bottle is on the floor.

VITR ✓ CLIP ✗

(a)

Query: A man is catching a yellow frisbee.

VITR ✓ CLIP ✗

(b)

Query: A white car in front of a city bus.

VITR ✓ CLIP ✗

(c)

Query: A girl is holding food.

VITR ✓ CLIP ✗

(d)

Fig. 7. A comparison of the top one results for text-to-image retrieval using CLIP and VITR. None of CLIP's results align with the description, whereas VITR's results are more relevant to the query description.

5.2 Visual Demonstrations of VITR's Advancements

The experiments on the datasets RefCOCOg and CLEVR, as discussed in Section 4.4, have quantitatively shown that VITR outperformed CLIP in relation-focused cross-modal information retrieval. This section provides examples to visually demonstrate the advancements that VITR has achieved in the tasks of image-to-text and text-to-image retrieval.

Image-to-Text Retrieval. Figure 6 presents eight examples of the top one image-to-text retrieval results between CLIP and VITR. Typically, CLIP's results offer a description of the image with errors or missing details of relations, while VITR's results concentrate on specific details. As seen in Figure 6(b), the result of CLIP describes the image as 'A mother zebra' without mentioning relations, while the result of VITR describes it as a relation-focused sentence which is 'A mother zebra nursing its baby'. Figure 6 highlights the limitations of CLIP in matching local image information, particularly relations, during image-to-text retrieval, and the improvement of VITR.

Text-to-Image Retrieval. Figure 7 presents four examples of the top one results of text-to-image retrieval between CLIP and VITR. As shown in Figure 7(c), the query aims to find an image of a white car in front of a bus, but the result from CLIP includes errors in the relations between the car and the bus, making the retrieved image less relevant to the query. On the other hand, VITR produces more accurate results that are better aligned with the intent of the query. Figure 7 highlights the limitations of CLIP in matching relation information between images and descriptions during text-to-image retrieval, and the improvement of VITR.

6 Conclusion

This article presents an innovative network that combines the local representations of an image with its global representation derived from the ViT model. The proposed network, VITR, is specifically designed for enhancing cross-modal information retrieval tasks. VITR includes a relational reasoning module that extends the capabilities of ViT by modeling the relations of regions in images for relation-focused cross-modal information retrieval; a fusion module that fuses the image global information from the ViT and the relation reasoned information of relational reasoning. Empirical evaluations revealed that the proposed VITR network outperformed CLIP and other VSE networks for both relation-focused and traditional cross-modal information retrieval tasks. When assessed through the average Recall@1 evaluation metric for retrieval performance, VITR exhibited superior results compared to CLIP. On the RefCOCOg dataset, VITR outperformed CLIP by 4.7% for image-to-text retrieval and 3.2% for text-to-image retrieval. On the CLEVR dataset, VITR achieved a substantial improvement of 24.2% for image-to-text retrieval and 18.5% for text-to-image retrieval. While VITR may not significantly outperform vision-language pre-trained models across all standard benchmarks, it shows notable improvements in relation-focused tasks. Additionally, VITR's unique relational reasoning and turbo modules enhance computational efficiency, making it particularly advantageous in applications requiring precise relation-focused retrieval and faster processing times. While the proposed VITR network is effective in image-to-text and text-to-image retrieval tasks, its limitation is that it does not consider other similar tasks, such as image captioning and Visual Question Answering. To overcome this limitation, future research could focus on extending VITR's capability to handle multiple tasks, thereby providing solutions for a broader range of applications. Furthermore, future work aims to integrate the capabilities of large language models into VITR and will conduct a comprehensive comparison with BLIP-2 [21].

References

[1] Jie Cao, Shengsheng Qian, Huaiwen Zhang, Quan Fang, and Changsheng Xu. 2021. Global relation-aware attention network for image-text retrieval. In *Proceedings of the International Conference on Multimedia Retrieval*, 19–28.

[2] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. 2021. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15789–15798.

[3] Ling Chen, Dandan Lyu, Shanshan Yu, and Gencai Chen. 2023. Multi-level visual similarity based personalized tourist attraction recommendation using geo-tagged photos. *ACM Transactions on Knowledge Discovery from Data* 17, 7 (2023), 1–18.

[4] Ran Chen, Hanli Wang, Lei Wang, and Sam Kwong. 2022. Two-stream hierarchical similarity reasoning for image-text matching. arXiv:2203.05349. Retrieved from https://doi.org/10.48550/arXiv.2203.05349

[5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: Universal image-text representation learning. In *Proceedings of the European Conference on Computer Vision,* 104–120.

[6] Mengjun Cheng, Yipeng Sun, Longchao Wang, Xiongwei Zhu, Kun Yao, Jie Chen, Guoli Song, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. 2022. ViSTA: Vision and scene text aggregation for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5184–5193.

[7] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1724–1734.

[8] Zhen Cui, Chunyan Xu, Wenming Zheng, and Jian Yang. 2018. Context-dependent diffusion network for visual relationship detection. In *Proceedings of the ACM International Conference on Multimedia*, 1475–1482.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 4171–4186.

[10] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. 2021a. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, 1218–1226.

[11] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference*, 12.

[12] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. StyleGAN-NADA: CLIP-guided domain adaptation of image generators. *ACM Transactions on Graphics* 41, 4 (2022), 1–13.

[13] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems* 33 (2020), 6616–6628.

[14] Yan Gong and Georgina Cosma. 2023. Improving visual-semantic embeddings by learning semantically-enhanced hard negatives for cross-modal information retrieval. *Pattern Recognition* 137 (2023), 109272.

[15] Yan Gong, Georgina Cosma, and Hui Fang. 2021. On the limitations of visual-semantic embedding networks for image-to-text information retrieval. *Journal of Imaging* 7, 8 (2021), 125.

[16] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2901–2910.

[17] Xiangyu Ke, Arijit Khan, and Francesco Bonchi. 2022. Multi-relation graph summarization. *ACM Transactions on Knowledge Discovery from Data* 16, 5 (2022), 1–30.

[18] John Boaz Lee, Ryan A Rossi, Sungchul Kim, Nesreen K Ahmed, and Eunyee Koh. 2019. Attention models in graphs: a survey. *ACM Transactions on Knowledge Discovery from Data* 13, 6 (2019), 1–25.

[19] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision*, 201–216.

[20] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, 11336–11344.

[21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv:2301.12597. Retrieved from https://dl.acm.org/doi/10.5555/3618408.3619222

[22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022b. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning.* PMLR, 12888–12900.

[23] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems* 34 (2021), 9694–9705.

[24] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019. Visual semantic reasoning for image-text matching. In *International Conference on Computer Vision (ICCV)*, 4654–4662.

[25] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2022c. Image-text embedding learning via visual and textual semantic reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 1 (2022), 641–656.

[26] Zejun Li, Zhihao Fan, Huaixiao Tou, Jingjing Chen, Zhongyu Wei, and Xuanjing Huang. 2022a. MVPTR: Multi-Level semantic alignment for vision-language pre-training via multi-stage learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4395–4405.

[27] Bingqian Lin, Yi Zhu, and Xiaodan Liang. 2022. Atom correlation based graph propagation for scene graph generation. *Pattern Recognition* 122 (2022), 108300.

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, 740–755.

[29] Fenglin Liu, Xian Wu, Shen Ge, Xuancheng Ren, Wei Fan, Xu Sun, and Yuexian Zou. 2021. DiMBERT: Learning vision-language grounded representations with disentangled multimodal-attention. *ACM Transactions on Knowledge Discovery from Data* 16, 1 (2021), 1–19.

[30] Haoyu Lu, Nanyi Fei, Yuqi Huo, Yizhao Gao, Zhiwu Lu, and Ji-Rong Wen. 2022. COTS: Collaborative two-stream vision-language pre-training model for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15692–15701.

[31] Haoyu Ma, Handong Zhao, Zhe Lin, Ajinkya Kale, Zhangyang Wang, Tong Yu, Jiuxiang Gu, Sunav Choudhary, and Xiaohui Xie. 2022b. EI-CLIP: Entity-aware interventional contrastive learning for e-commerce cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18051–18061.

[32] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. 2022a. X-CLIP: end-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the ACM International Conference on Multimedia*, 638–647.

[33] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11–20.

[34] Mingyuan Mao, Peng Gao, Renrui Zhang, Honghui Zheng, Teli Ma, Yan Peng, Errui Ding, Baochang Zhang, Shumin Han. 2021. Dual-stream network for visual recognition. *Advances in Neural Information Processing Systems* 34 (2021), 25346–25358.

[35] Sunil Kumar Maurya, Xin Liu, and Tsuyoshi Murata. 2021. Graph neural networks for fast node ranking approximation. *ACM Transactions on Knowledge Discovery from Data* 15, 5 (2021), 1–32.

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* (*ICML)*, 8748–8763.

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 6 (2016), 1137–1149.

[38] Ming Tao, Bing-Kun Bao, Hao Tang, and Changsheng Xu. 2023. GALIP: Generative adversarial CLIPs for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14214–14223.

[39] Enqiang Wang, Qing Yu, Yelin Chen, Wushouer Slamu, and Xukang Luo. 2022. Multi-modal knowledge graphs representation learning via multi-headed self-attention. *Information Fusion* 88 (2022), 78–85.

[40] Jianyu Wang, Bing-Kun Bao, and Changsheng Xu. 2021. DualVGR: A dual-visual graph reasoning unit for video question answering. *IEEE Transactions on Multimedia* 24 (2021), 3369–3380.

[41] Yan Wang, Yuting Su, Wenhui Li, Zhengya Sun, Zhiqiang Wei, Jie Nie, Xuanya Li, and An-An Liu. 2023. Rare-aware attention network for image–text matching. *Information Processing & Management* 60, 3 (2023), 103280.

[42] Yaxiong Wang, Hao Yang, Xiuxiu Bai, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. 2020. PFAN++: Bi-directional image-text retrieval with position focused attention network. *IEEE Transactions on Multimedia* 23 (2020), 3362–3376.

[43] Ying Wei, Yangqiu Song, Yi Zhen, Bo Liu, and Qiang Yang. 2016. Heterogeneous translated hashing: A scalable solution towards multi-modal similarity search. *ACM Transactions on Knowledge Discovery from Data* 10, 4 (2016), 1–28.

[44] Ke Yan, Yaowei Wang, Dawei Liang, Tiejun Huang, and Yonghong Tian. 2016. CNN vs. SIFT for image retrieval: Alternative or complementary?. In *Proceedings of the ACM International Conference on Multimedia*, 407–411.

[45] Sangwoong Yoon, Woo Young Kang, Sungwook Jeon, SeongEun Lee, Changjin Han, Jonghun Park, and Eun-Sol Kim. 2021. Image-to-image retrieval by learning similarity between scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, 10718–10726.

[46] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2 (2014), 67–78.

[47] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE-ViL: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, 3208–3216.

[48] Lisai Zhang, Qingcai Chen, Zhijian Chen, Yunpeng Han, Zhonghua Li, and Zhao Cao. 2023a. Refined vision-language modeling for fine-grained multi-modal pre-training. arXiv:2303.05313. Retrieved from https://doi.org/10.48550/arXiv.2303.05313

[49] Si Zhang, Hanghang Tong, Jiejun Xu, and Ross Maciejewski. 2019. Graph convolutional networks: A comprehensive review. *Computational Social Networks* 6, 1 (2019), 1–23.

[50] Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. 2023c. Boosting scene graph generation with visual relation saliency. *ACM Transactions on Multimedia Computing, Communications and Applications* 19, 1 (2023), 1–17.

[51] Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2023b. Universal multimodal representation for language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1 (2023), 1–18.

[52] Zheng Zhang, Xiaofeng Zhu, Guangming Lu, and Yudong Zhang. 2021. Probability ordinal-preserving semantic hashing for large-scale image retrieval. *ACM Transactions on Knowledge Discovery from Data* 15, 3 (2021), 1–22.

Received 28 June 2023; revised 16 May 2024; accepted 28 July 2024