Applied Intelligence (2023) 53:22898–22916
https://doi.org/10.1007/s10489-023-04669-3

TIAR: Text-Image-Audio Retrieval with weighted multimodal re-ranking

Peide Chi[1] · Yong Feng[1] · Mingliang Zhou[1] · Xian-cai Xiong[2,3] · Yong-heng Wang[4] · Bao-hua Qiang[5]

Accepted: 25 April 2023 / Published online: 4 July 2023

Abstract

Cross-modal retrieval has developed remarkably recently and received extensive attention as an essential method for multimodal interaction study. However, most existing models are limited to one of the applications in cross-modal retrieval, i.e., text-image retrieval, and neglect the audio modality, which is widely distributed in data and can be integrated into the models to improve retrieval performance. To address this issue, we propose a text-image-audio cross-modal retrieval (TIAR) model that, given any or two modalities, implements the retrieval of the remaining modalities. TIAR consists of three modal-specific encoders to extract the features and a cross-modal encoder to generate joint contextualized representations for all modalities. To evaluate our model, we present two new cross-modal retrieval tasks, named cross-unimodal and cross-bimodal retrieval, that are applicable to three modalities. Then, during testing, we propose a weighted multimodal re-ranking (WMR) algorithm which integrates comprehensive ranking information in the similarity matrices of all tasks to improve the performance without additional training. The experiment results show that TIAR-WMR outperforms state-of-the-art models in traditional text-image retrieval on Flickr30k, COCO, and ADE20k datasets. Moreover, the retrieval performance of TIAR-WMR is further boosted in the two proposed tasks when two input modalities are integrated. The code is available at https://github.com/PeideChi/TIAR.

Keywords Cross-modal retrieval · Audio retrieval · Multimedia · Fusion learning

1 Introduction

In modern society, due to the rapid development of technology, human beings live in a world full of data. Data has diverse sources and applications and can be presented in different modalities. How to effectively process these enormous amounts of technological by-products has been a long-standing problem for researchers. Cross-modal retrieval has received considerable attention as one of the approaches to solving the problem of processing data with the same semantics but different modalities.

The past few years have seen increasingly rapid advances in large pre-trained models in natural language processing [11, 58] and computer vision [16, 34]. Numerous works [7, 8, 14, 20, 44] have been published on cross-modal retrieval based on these pre-trained models. Extensive researches have been carried out on text-image retrieval, one of the tasks of cross-modal retrieval, but few studies investigate those retrieval tasks that comprehensively process three modalities. We are interested in audio, a modality widely distributed in data, and considering integrating it into text-image retrieval. Both text and image modalities correspond to only the visual mode in human sensory modes, while the audio modality corresponds to the auditory mode that ought to have equal importance for humans as the former [49]. All sensory modes are helpful and synthetically contribute to our awareness of the environment and understanding of the world [50]. It is a pity that there are few models that can conduct cross-retrieval among the three modalities of text, image, and audio.

Besides, automatic speech recognition (ASR) has been widely studied by computer scientists over the past several decades. Speech is the most efficient, preferred, and natural way for humans to communicate with each other. ASRs are considered to be the future means of communication between humans and machines [38]. Therefore, audio is a significant modality that should be considered in cross-modal retrieval. Using audio for retrieval or retrieving audio is an essential retrieval task. Integrating the audio modality into the model

✉ Yong Feng
fengyong@cqu.edu.cn

Extended author information available on the last page of the article

can improve retrieval performance and make cross-modal retrieval more generalized.

In this paper, we propose a **T**ext-**I**mage-**A**udio cross-modal **R**etrieval (TIAR) model to perform a more comprehensive retrieval task. Concretely, TIAR consists of three modal-specific encoders for text, image, and audio respectively and a cross-modal encoder that conducts cross-attention among the three modalities to learn cross-modal alignment. We input texts, images, and audios to the corresponding modal-specific encoders of TIAR to generate modal features at different semantic levels. Then we gather all features of the three modalities and learn joint contextualized representation for each modality through the cross-modal encoder. To better align the semantics in different modalities and learn cross-modal interaction, all modules in TIAR utilize a full transformer design. Benefiting from the excellent performance of the attention mechanism, TIAR works very well in multimodal fusion. During testing, TIAR inputs one or two modalities and returns the remaining modalities that are most similar to the input modalities, corresponding to two new proposed retrieval tasks, cross-unimodal and cross-bimodal retrieval.

In addition, multimodal fusion, an essential learning scheme in training, is ignored by most of the previous retrieval models during testing, which results in a discrepancy between training and testing. To integrate this scheme into the testing phase, Wang et al. [54] came up with a basic assumption: if one modality X has a high ranking among the retrieval candidates of another modality Y, Y is also in front of the candidates of X. Following this assumption, we propose a weighted multimodal re-ranking (WMR) algorithm. By using the top K retrieval candidates to perform a reverse search, WMR fuses the ranking information of similarity matrices of multimodal into the original ranking results. Our algorithm bridges the gap between training and testing and remarkably improves retrieval performance without extra training procedures.

Experimental results demonstrate that TIAR-WMR achieves state-of-the-art performance in traditional text-image retrieval and shows promising results in the two proposed retrieval tasks on Flickr30k, COCO, and ADE20k datasets. Moreover, experimental results also prove that the retrieval performance of TIAR-WMR is further improved on the three benchmarks when two input modalities are integrated, demonstrating our model's impressive multimodal fusion capability.

The main contributions of this work are summarized as follows:

- We propose a new cross-modal retrieval model, TIAR, that takes the text, image, and audio data as inputs and implements cross-retrieval among the three modalities. Two new cross-modal retrieval tasks, named cross-unimodal and cross-bimodal, are presented to evaluate the retrieval performance of TIAR.
- We propose a weighted multimodal re-ranking (WMR) algorithm that makes full use of the ranking information of similarity matrices in all tasks to improve retrieval accuracy without additional training.
- TIAR-WMR outperforms state-of-the-art methods remarkably in traditional text-image retrieval on Flickr30k, COCO, and ADE20k datasets. Moreover, the performance is further boosted on three experimental benchmarks when two input modalities are integrated.

The rest of this paper is organized as follows. First the related works about our work are reviewed in Section 2. The specific implementation details of our method are described in Section 3. Finally, we analyze the experiments and summarize our work, outlook for future work respectively in Sections 4 and 5.

2 Related works

2.1 Text-image retrieval

The existing research on cross-modal retrieval is extensive and focuses mainly on one of the applications, text-image retrieval. Text-image retrieval aims to obtain the most relevant images or text descriptions given a query text or image. The text-image retrieval model consists of two main components: the embedding of text and image inputs and the multimodal fusion in the deep network.

Text and Image Embedding As the first module of the text-image retrieval model, embedding is used to map discrete inputs of different modalities to a uniform dimensional space to facilitate the data processing of the model. For text modality, some models [13, 62, 66] mainly use RNNs to learn the representation of sentences. Since the emergence of attention mechanism [52], transformers, especially BERT [24], have led many later proposed models [22, 32, 64, 68] to adopt them as the text encoders of these models with their superior global feature extraction capability over RNNs. These models achieve a considerable performance improvement benefiting from the text processing ability of pre-trained BERT. For image modality, there are two common approaches to embedding images. The first approach is to extract the region features of the image, which are obtained by an object detector like Faster R-CNN [45] and are also referred to as bottom-up features [1]. The retrieval models using this method include IMRAM [4], SGRAF [12], and L3S-KD [63]. These region features are usually extracted offline and stored as a specific file instead of the original images, which has the advantage that the extraction process

can be done in advance without occupying model training and testing time. It should be noted that this approach is hard to reduce the time spent when processing brand-new images that do not exist in the dataset. The second approach introduced in [21] is to feed the image into a convolutional neural network such as ResNet [15] or other image transformers like Swin-Transformer [34] and return the output image grid features. The primary motivation for this approach is to avoid the slow extraction process of the object detector and its possible extraction errors. Some cross-modal retrieval models based on this approach, such as X-LXMERT [8], X-VLM [60], and ViSTA [7], have been proposed successively and have shown promising retrieval results.

Multimodal Fusion The multimodal fusion module is the core of the text-image retrieval model. Its function is to semantically align the extracted input modality embeddings in a uniform high-dimensional space and learn joint contextualized representations for all feature embeddings in the input sequence. A typical but simple multimodal fusion approach is to directly dot product the feature embeddings of the input modalities or feed them into several shallow layers. Several models based on this approach, such as CRGN [65], CLIP [44], and MLMUG [37], demonstrate competitive retrieval accuracy. As proved in [26], the simple multimodal fusion approach may perform fairly well when the size of the dataset is not too large, yet it is not sufficient for large-scale datasets or tasks with high modal-fusion requirements. Recently, an increasing number of models using deep multimodal fusion modules have been proposed, such as ALIGN [20], TCL [57], and METER [14], which are usually stacked by several transformer layers. More sophisticated multimodal fusion modules boost the retrieval performance of the models, but they also increase the complexity of these models and require more time and data to train them. That is a tradeoff between performance and efficiency that should be considered for all models when adapting a multimodal fusion scheme.

2.2 Pre-training models

As a mainstream paradigm in computer vision, natural language processing, and other research areas, pre-training-then-fine-tuning remarkably improves model performance in a number of downstream tasks. Recently, visual language pre-training models have also developed significantly, especially in tasks such as cross-modal retrieval and visual question answering.

Single-Modality Pre-Training In computer vision, numerous methods adopt pre-training for multimodal tasks. ResNet and Faster R-CNN mentioned above are two of the most classic pre-trained visual feature extractors. With the introduction of transformer, an increasing number of pre-trained visual models have been proposed, and their performance in conventional tasks such as image classification, semantic segmentation, and object detection has been improved. Some current state-of-the-art models include NFNet [3], MFF-PCB [53], KAZSLM [27], and ViT-G [61]. In natural language processing, the introduction of transformer has also led to a promising advancement in pre-trained language models. Compared to RNN, transformer focuses more on the global feature representations of all words in a sentence and can process the whole sentence at once. It is due to these advantages that transformer-based methods are leading the way in various downstream tasks and have become the dominant and most preferred paradigm in the field. Among these language pre-training models, BERT is the most used models. Besides, some representative latest models include Routing Transformer [46] and LTFE [51] . For audio processing, an increasing number of current mainstream models are gradually inspired by transformer in several downstream tasks. Taking speech recognition as an example, wav2vec2.0 [2] introduces transformer on the basis of wav2vec [47], which reduces the word error rate of the model and becomes the basis of successive models [9, 56]. In addition, pre-trained models also contribute to other tasks, such as emotion recognition [35, 36], speech synthesis [5, 55], and spoken language understanding [25, 43, 48].

Multi-Modal Pre-Training Pre-trained multimodal models perform impressively on many visual and language tasks based on transformer. These models can be divided into two main categories: single-encoder and dual-encoder. The single-encoder models [6, 19, 33] use one multimodal transformer to fuse image and text features for modal interaction. Although this method has excellent performance on some downstream tasks, its computational cost is too high to be applied to large-scale cross-modal datasets. To cope with this drawback, the dual-encoder models [17, 23, 42] construct two separate encoders for images and texts, significantly reducing the computation time of the similarity of the image-text pairs.

In contrast with the models mentioned above, TIAR considers three input modalities simultaneously, i.e., text, image, and audio, and conducts cross-modal retrieval for these three modalities given any one or two input modalities.

3 Proposed method

Figure 1 illustrates the architecture of the proposed model TIAR and the weighted multimodal re-ranking algorithm. TIAR consists of two parts, i.e., three modal-specific encoders and one cross-modal encoder. We adopt a full transformer design for the three modal-specific encoders, namely text, image, and audio encoder, to generate effective feature representations for higher retrieval accuracy. After encoding the inputs of the three modalities, we gather all the embeddings as a sequence and feed them into the cross-modal encoder

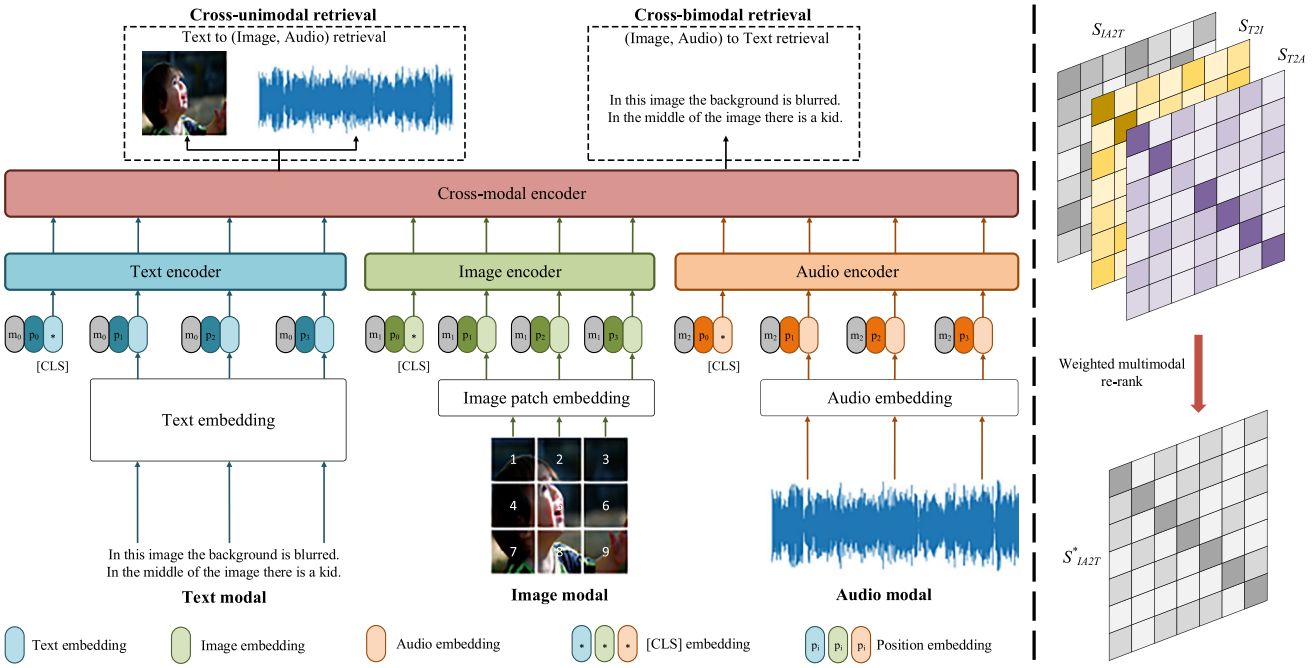


Fig. 1 Architecture of the proposed model TIAR (left) and weighted multimodal re-rank (right)

for multimodal fusion. To better fuse the three modalities, the cross-modal encoder conducts cross-attention to all the embeddings in the sequence. Contrastive learning loss and matching prediction loss are used as the training loss function of the model, encouraging the model to align the semantics of embeddings of different modalities and generate feature representations that fuse multimodal information. We finally use the three [CLS] token embeddings corresponding to the three modalities for two new proposed cross-modal retrieval tasks, named cross-unimodal and cross-bimodal retrieval.

Let $X = \{(T_i, I_i, A_i)\}_{i=1}^{N}$ be the dataset, where (T_i, I_i, A_i) are the text, image, and audio for the ith sample, and N is the size of the dataset. Denote $\mathbb{T} = \{T_i\}_{i=1}^{N}$, $\mathbb{I} = \{I_i\}_{i=1}^{N}$, and $\mathbb{A} = \{A_i\}_{i=1}^{N}$ as the text, image, and audio set in X respectively. Depending on the number of input modalities, the object of cross-modal retrieval is to conduct two tasks (the details are defined in Section 3.3):

- **Cross-Unimodal Retrieval** Given one sample of any modality, find two samples in the dataset corresponding to each of the two remaining modalities that are most similar to it respectively.
- **Cross-Bimodal Retrieval** Given two samples of any two modalities, find one sample in the dataset corresponding to the remaining modality that is most similar to them simultaneously.

3.1 Modal-specific encoder

Text Encoder The text encoder stacks of several standard transformer layers from BERT. We first tokenize all words in the caption to obtain the token sequence $T = [t_1, t_2, ..., t_n]$, where t_i is the ith word token and n is the length of the token sequence. Let $T^0 = [t^0_{\text{cls}}, t_1, t_2, ..., t_n] = [t^0_{\text{cls}}, t^0_1, t^0_2, ..., t^0_n]$ be all of input word tokens concatenated with a [CLS] token t^0_{cls}. By feeding T^0 into the text encoder, we obtain the embeddings of all words and the [CLS] token of the input caption. The process can be expressed by the following equations:

$$\begin{aligned} \hat{T}^l &= \text{MHSA}(\text{LN}(T^{l-1})) \\ T^l &= \text{MLP}(\text{LN}(\hat{T}^l)) \end{aligned} \tag{1}$$

where $\hat{T}^l$ and T^l denote the output embeddings of MHSA and MLP for layer $l (l = 1, 2, ..., N_T)$, MHSA denotes multi-head self-attention, MLP denotes multi-layer perception, LN denotes the layer normalization, and N_T is the number of transformer layers in the text encoder. The final embeddings for all tokens T^E is obtained by passing through a layer normalization layer $T^E = \text{LN}(T^{N_T}) = \text{LN}([t^{N_T}_{\text{cls}}, t^{N_T}_1, t^{N_T}_2, ..., t^{N_T}_n])$.

The text encoder takes the original text caption T as input and outputs the text embeddings T^E, where each $t^E \in \mathbb{R}^{d_T}$ and $d_T = 768$ in the configuration of BERT-base. Different

from BERT-base, to reduce the number of parameters of the model, the text encoder uses only the first six transformer layers, i.e., $N_T = 6$. The configuration of the text encoder is inspired by [60]. Experiments show that this setting is adequate to make the model perform competitively in text-related tasks and is a well-balanced tradeoff between performance and efficiency.

Image Encoder The image encoder efficiently generates both global and multi-grained vision representations of an image. Following Swin-Transformer [34], we first split an input RGB image I of resolution of 224×224 into non-overlapping patches of size of 32×32. Each patch is treated as a "token", and its feature is the concatenation of the raw pixel RGB values.

Denote the input image and its patches as $I = [p_1, p_2, ..., p_{49}] \in \mathbb{I} = \mathbb{R}^{3\times224\times224}$, where each $p_i (i = 1, 2, ..., 49) \in \mathbb{R}^{3\times32\times32}$. We first feed all patches into a linear patch embedding (LPE) layer and then concatenate with a [CLS] token to obtain the token sequence of the input image $I^0 = [p^0_{\text{cls}}, \text{LPE}(I)] = [p^0_{\text{cls}}, p^0_1, p^0_2, ..., p^0_{49}]$. We calculate the embeddings of all patches in the image as follows:

$$\begin{aligned} \hat{I}^l &= \text{WA}(\text{LN}(I^{l-1})) \\ I^l &= \text{MLP}(\text{LN}(\hat{I}^l)) \end{aligned} \tag{2}$$

where $\hat{I}^l$ and I^l denote the output embeddings of WA and MLP for layer $l (l = 1, 2, ..., N_I)$, WA denotes window attention, and N_I is the number of layers in the image encoder. After that, we obtain all the embeddings $I^{N_I} = [p^{N_I}_{\text{cls}}, p^{N_I}_1, p^{N_I}_2, ..., p^{N_I}_{49}]$. We pass I^{N_I} through a layer normalization layer to calculate the final embeddings of tokens $I^E = \text{LN}(I^{N_I})$.

The image encoder inputs the original image I and outputs the image embeddings I^E, where each $p^E \in \mathbb{R}^{d_I}$ and $d_I = 1024$. By utilizing both global and multi-grained information of the image, the image embeddings are sufficient to represent the image in the subsequent processing.

Audio Encoder We use wav2vec2.0 [2] as the audio encoder. Given the raw audio A, we initially feed it into a convolutional feature encoder to get the latent representations of the input audio $\text{CNN}(A) = [a^0_1, a^0_2, ..., a^0_m]$, where m is the length of audio representations. And then, we input these representations concatenated with a [CLS] token $A^0 = [a^0_{\text{cls}}, a^0_1, a^0_2, ..., a^0_m]$ into the audio encoder to calculate their embedding as follows:

$$\begin{aligned} \hat{A}^l &= \text{MHSA}(\text{LN}(A^{l-1})) \\ A^l &= \text{MLP}(\text{LN}(\hat{A}^l)) \end{aligned} \tag{3}$$

where $\hat{A}^l$ and A^l denote the output embeddings of MHSA and MLP for layer $l (l = 1, 2, ..., N_A)$, and N_A is the number of transformer layers in the audio encoder. As above, we also pass A^{N_A} through a layer normalization layer to obtain the final embeddings of audio $A^E = \text{LN}(A^{N_A})$.

The audio encoder takes the raw audio A as input and outputs the audio embeddings A^E, where each $a^E \in \mathbb{R}^{d_A}$ and $d_A = 1024$. TIAR benefits from this model and performs surprisingly well in audio-involved tasks after carefully adapting and training.

3.2 Cross-modal encoder

The extraction of text, image, and audio embeddings is isolated from cross-modal interaction. In order to fuse the embeddings obtained by independent encoders, the cross-attention mechanism is used to extract effective features and semantic information. We gather all embeddings of the three modalities as a sequence $Z^E = [T^E; I^E; A^E] = [t^E_{\text{cls}}, t^E_1, ..., t^E_n; p^E_{\text{cls}}, p^E_1, ..., p^E_{49}; a^E_{\text{cls}}, a^E_1, ..., a^E_m]$. Since the dimensions of the embeddings of the three modalities are not uniform, we first pass a multi-layer perception layer to align the dimensions of text embeddings to 1024 before inputting them into the cross-modal encoder $Z^0 = [\text{MLP}(T^E); I^E; A^E]$. Then, we feed the aligned embedding sequence Z^0 into the cross-modal encoder. The cross-modal encoder processes all embeddings of the three modalities as follows:

$$\begin{aligned} \hat{Z}^l &= \text{CA}(\text{LN}(Z^{l-1})) \\ Z^l &= \text{MLP}(\text{LN}(\hat{Z}^l)) \end{aligned} \tag{4}$$

where $\hat{Z}^l$ and Z^l denote the output embeddings of CA and MLP for layer $l (l = 1, 2, ..., N_C)$, CA denotes cross-attention and N_C is the number of layers in the cross-modal encoder. We experiment N_C with 3 values which are 6, 7, 8, and set $N_C = 7$ for the best performance (see details in Section 4.4.2).

After multiple layers of cross-attention processing, the [CLS] token embeddings of three modalities $t^{N_C}_{\text{cls}}, p^{N_C}_{\text{cls}}, a^{N_C}_{\text{cls}} \in Z^{N_C}$ fuse the semantics among all modalities which can be used for cross-modal retrieval.

3.3 Cross-modal retrieval

Depending on the number of input modalities, the cross-modal retrieval has the following two tasks, namely cross-unimodal and cross-bimodal retrieval.

Cross-Unimodal Retrieval Figure 2 illustrates an example of cross-unimodal retrieval. Given one sample of any modality, without loss of generality, let it be $T \in \mathbb{T}$, the goal function g_1 calculates the similarity between T and all samples of image and audio modalities in the dataset and selects

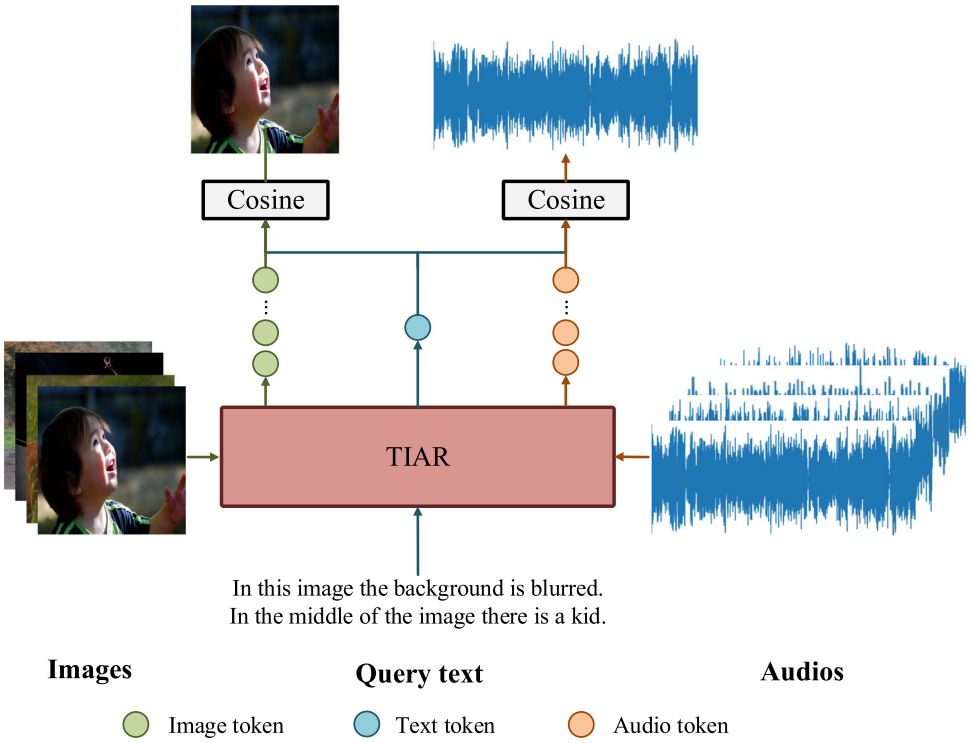


Fig. 2 An example of the proposed cross-unimodal retrieval. Given a query text T, the aim is to obtain an image and audio in the dataset, which are the most similar to T respectively

one sample from each modality that is most similar to T as follows:

$$\mathrm{g}_1(T) = (\arg\max_{I\in\mathbb{I}}\{\mathrm{S}(\mathrm{h}(I),\mathrm{h}(T))\}, \arg\max_{A\in\mathbb{A}}\{\mathrm{S}(\mathrm{h}(A),\mathrm{h}(T))\}) \tag{5}$$

where $\mathrm{S}(\cdot,\cdot)$ is the cosine similarity measurement function, and $\mathrm{h}(\cdot)$ is the function that maps the original input of each modality to the output [CLS] token embedding of the cross-modal encoder in Eq. 4:

$$\mathrm{h}(X) = \begin{cases} t_{\mathrm{cls}}^{N_C}, & X = T \\ p_{\mathrm{cls}}^{N_C}, & X = I \\ a_{\mathrm{cls}}^{N_C}, & X = A \end{cases} \tag{6}$$

Similarly, we define the function g_1 for the cases of image and audio modalities in the same manner. The overall equation of g_1 is defined as follows:

$$\mathrm{g}_1(X) = \begin{cases} (\arg\max_{I\in\mathbb{I}}\{\mathrm{S}(\mathrm{h}(I),\mathrm{h}(T))\}, \arg\max_{A\in\mathbb{A}}\{\mathrm{S}(\mathrm{h}(A),\mathrm{h}(T))\}), & X = T \\ (\arg\max_{T\in\mathbb{T}}\{\mathrm{S}(\mathrm{h}(T),\mathrm{h}(I))\}, \arg\max_{A\in\mathbb{A}}\{\mathrm{S}(\mathrm{h}(A),\mathrm{h}(I))\}), & X = I \\ (\arg\max_{T\in\mathbb{T}}\{\mathrm{S}(\mathrm{h}(T),\mathrm{h}(A))\}, \arg\max_{I\in\mathbb{I}}\{\mathrm{S}(\mathrm{h}(I),\mathrm{h}(A))\}), & X = A \end{cases} \tag{7}$$

Cross-Bimodal Retrieval Figure 3 illustrates an example of cross-bimodal retrieval. Given two samples of any two modalities, without loss of generality, let them be $I \in \mathbb{I}$ and $A \in \mathbb{A}$, the goal function g_2 calculates the similarity between I and A and all samples of text modality in the dataset and selects one sample from them that is most similar to both I and A as follows:

$$\mathrm{g}_2(I, A) = \arg\max_{T\in\mathbb{T}}\{\alpha\,\mathrm{S}(\mathrm{h}(T),\mathrm{h}(I)) + (1-\alpha)\,\mathrm{S}(\mathrm{h}(T),\mathrm{h}(A))\} \tag{8}$$

where α is a hyper-parameter adjusting the combination of the two modalities.

Similarly, for the other two cases, we define the function g_2 in the same manner. The overall equation of g_2 is defined as follows:

$$\mathrm{g}_2(X, Y) = \begin{cases} \arg\max_{T\in\mathbb{T}}\{\alpha\,\mathrm{S}(\mathrm{h}(T),\mathrm{h}(I)) + (1-\alpha)\,\mathrm{S}(\mathrm{h}(T),\mathrm{h}(A))\}, & (X,Y) = (I,A) \\ \arg\max_{A\in\mathbb{A}}\{\beta\,\mathrm{S}(\mathrm{h}(A),\mathrm{h}(T)) + (1-\beta)\,\mathrm{S}(\mathrm{h}(A),\mathrm{h}(I))\}, & (X,Y) = (T,I) \\ \arg\max_{I\in\mathbb{I}}\{\gamma\,\mathrm{S}(\mathrm{h}(I),\mathrm{h}(T)) + (1-\gamma)\,\mathrm{S}(\mathrm{h}(I),\mathrm{h}(A))\}, & (X,Y) = (T,A) \end{cases} \tag{9}$$

where β and γ are also hyper-parameters.

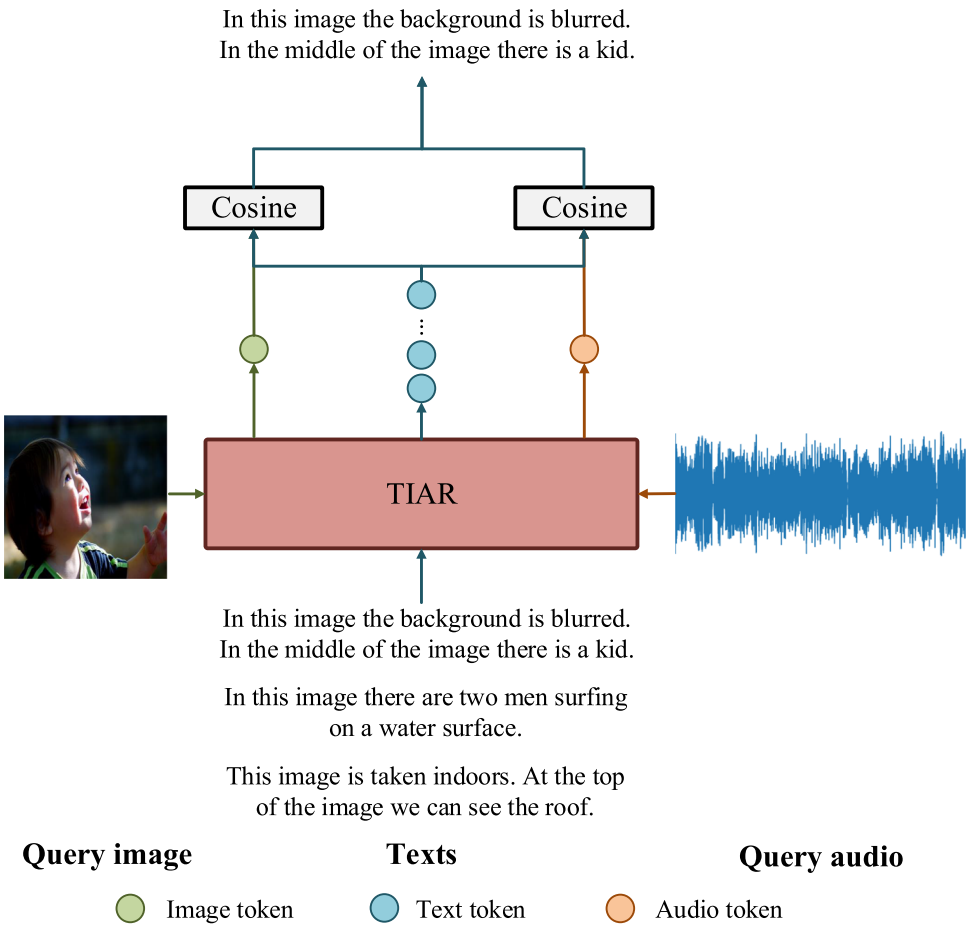


Fig. 3 An example of the proposed cross-bimodal retrieval. Given a query image I and audio A, the aim is to obtain a text which is the most similar to I and A simultaneously in the dataset

3.4 Loss function

Contrastive Learning Loss Contrastive learning loss encourages the model to distinguish samples at semantic level and focus on learning common features among similar samples. We sample a random mini-batch of m samples $b_m = \{(T_i, I_i, A_i)\}_{i=1}^{m}$, and calculate the in-batch cross-modal similarities.

Concretely, for any sample $(T_i, I_i, A_i) \in b_m$ and any two of its modalities, without loss of generality, let them be T_i and I_i. We treat I_i as the positive example for T_i and the rest of the $m-1$ images within b_m as the negative examples, and vice versa. Then, we calculate T_i's text-to-image similarity to each image I_j and I_i's image-to-text similarity to each text T_j in the mini-batch as:

$$s_i^{\mathrm{T2I}}(I_j) = \frac{\exp(\mathrm{S}(\mathrm{p}(T_i), \mathrm{p}(I_j))/\tau)}{\sum_{k=1}^{m} \exp(\mathrm{S}(\mathrm{p}(T_i), \mathrm{p}(I_k))/\tau)} \tag{10}$$

$$s_i^{\mathrm{I2T}}(T_j) = \frac{\exp(\mathrm{S}(\mathrm{p}(T_j), \mathrm{p}(I_i))/\tau)}{\sum_{k=1}^{m} \exp(\mathrm{S}(\mathrm{p}(T_k), \mathrm{p}(I_i))/\tau)} \tag{11}$$

where τ is a learnable temperature parameter, and $\mathrm{p}(\cdot)$ is the transformation that map the original input of each modality to the normalized output [CLS] token embedding of its corresponding modal-specific encoder:

$$\mathrm{p}(X) = \begin{cases} t_{\mathrm{cls}}^{E}, & X = T \\ p_{\mathrm{cls}}^{E}, & X = I \\ a_{\mathrm{cls}}^{E}, & X = A \end{cases} \tag{12}$$

Let $\boldsymbol{y}_i^{\mathrm{T2I}}$ and $\boldsymbol{y}_i^{\mathrm{I2T}}$ be the ground-truth one-hot similarity vectors of T_i and I_i in the mini-batch, where only the positive example has the probability of one, and the remaining negative examples have zero. $\boldsymbol{s}_i^{\mathrm{T2I}}$ and $\boldsymbol{s}_i^{\mathrm{I2T}}$ are the text-to-image and image-to-text similarity vectors of T_i and I_i in the mini-batch. The text-image contrastive learning loss of b_m is defined as follows:

$$\mathcal{L}_{\mathrm{cl}}^{\mathrm{TI}} = \frac{1}{2}\left(\frac{\sum_{i=1}^{m} \mathrm{H}(\boldsymbol{y}_i^{\mathrm{T2I}}, \boldsymbol{s}_i^{\mathrm{T2I}})}{m} + \frac{\sum_{i=1}^{m} \mathrm{H}(\boldsymbol{y}_i^{\mathrm{I2T}}, \boldsymbol{s}_i^{\mathrm{I2T}})}{m}\right) \tag{13}$$

where $\mathrm{H}(\boldsymbol{p}_1, \boldsymbol{p}_2)$ is the cross-entropy of two distributions $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$. In the same manner, we can define text-audio loss $\mathcal{L}_{\mathrm{cl}}^{\mathrm{TA}}$ and image-audio loss $\mathcal{L}_{\mathrm{cl}}^{\mathrm{IA}}$ in this mini-batch. The total

contrastive learning loss in the mini-batch is defined as:

$$\mathcal{L}_{\text{cl}} = \frac{1}{3}(\mathcal{L}_{\text{cl}}^{\text{TI}} + \mathcal{L}_{\text{cl}}^{\text{TA}} + \mathcal{L}_{\text{cl}}^{\text{IA}}) \tag{14}$$

Matching Prediction Loss Matching prediction loss encourages the model to match samples in the dataset as many as possible. Inspired by [60], we sample one hard negative image and audio respectively by following $\boldsymbol{s}^{\text{T2I}}$ and $\boldsymbol{s}^{\text{T2A}}$ for each text in the mini-batch b_m. For image and audio modalities, we perform the same sampling procedure as above. The more similar the two examples are to each other, the higher the probability they will be sampled.

Concretely, for any example $X_i \in \{T_i, I_i, A_i\}$ and its hard negative example, without loss of generality, let $X_i = T_i$, and $I_j (j = 1, 2, ..., m)$ be its hard negative example. we calculate the matching probability $\boldsymbol{p}_i^{\text{T2I}}$ by feeding the output [CLS] embedding of the cross-modal encoder into a specific multi-layer perception layer. The text-to-image matching prediction loss of b_m is the mean value of the cross-entropy H between $\boldsymbol{p}_i^{\text{T2I}}$ and $\boldsymbol{y}_i^{\text{T2I}}$:

$$\mathcal{L}_{\text{match}}^{\text{T2I}} = \frac{\sum_{i=1}^{m} \text{H}(\boldsymbol{y}_i^{\text{T2I}}, \boldsymbol{p}_i^{\text{T2I}})}{m} \tag{15}$$

where $\boldsymbol{y}_i^{\text{T2I}}$ is a 2-dimensional one-hot vector representing the ground-truth label. We can define the image-to-text matching prediction loss of b_m in the same manner, and formulate the text-image loss as:

$$\mathcal{L}_{\text{match}}^{\text{TI}} = \frac{1}{2}(\mathcal{L}_{\text{match}}^{\text{T2I}} + \mathcal{L}_{\text{match}}^{\text{I2T}}) \tag{16}$$

The total matching prediction loss of b_m is defined as:

$$\mathcal{L}_{\text{match}} = \frac{1}{3}(\mathcal{L}_{\text{match}}^{\text{TI}} + \mathcal{L}_{\text{match}}^{\text{TA}} + \mathcal{L}_{\text{match}}^{\text{IA}}) \tag{17}$$

Finally, the overall loss of our proposed TIAR can be formulated as follows:

$$\mathcal{L} = \mathcal{L}_{\text{cl}} + \mathcal{L}_{\text{match}} \tag{18}$$

3.5 Weighted multimodal re-rank

The retrieval results are obtained by using only the corresponding similarity matrix for each task, for example, the similarity S^{I2T} for image-to-text retrieval. Wang et al. [54] argued that low testing accuracy is found due to ignoring the interactions between text-to-image and image-to-text retrieval and proposed a cross-modal re-ranking method. Although the performance is improved, the cross-modal re-ranking considers only two reverse similarity matrices, which are still insufficient for both the cross-unimodal and cross-bimodal retrieval.

In order to fully integrate the information in similarity matrices of all modalities, we propose a weighted multimodal re-rank (WMR) algorithm, shown in Fig. 4. Following the basic assumption that any sample in a triplet (T, I, A) is supposed to be retrieved by the rest in all retrievals forwardly and backwardly, we utilize the similarity matrices, which is reverse to the retrieval, to correct errors made from the original similarity matrix and improve the retrieval accuracy. After obtaining the original similarity matrices, we select the K-nearest neighbors for each retrieved candidate and reversely search the rank of the query for all of the K candidates. We then use a ranking position set to calculate an importance vector and modify the similarity between the query and each candidate. For cross-unimodal and cross-bimodal retrieval, we propose two weighted multimodal re-ranking strategies.

Cross-Unimodal Retrieval We take the image-to-text retrieval task as an example to explain the details of WMR which is shown in Algorithm 1. Given a query image I_q and its initial vector $\boldsymbol{s}_q^{\text{I2T}}$ in similarity matrix S^{I2T}, we select K-nearest neighbor texts which have the top K maximum similarity to I_q and denote them as $R^{\text{I2T}}(I_q, K) = \{T_1^q, T_2^q, ..., T_K^q\}$, where K is the number of the nearest neighbors. Then for each text $T_j^q \in R^{\text{I2T}}(I_q, K)$, the reverse search process is performed. Concretely, we define the K-nearest images of T_j^q as $R^{\text{T2I}}(T_j^q, K) = \{I_1^j, I_2^j, ..., I_K^j\}$, where I_i^j is the image which has the ith largest similarity to T_j^q according to S^{T2I}. To integrate the similarity information in both nearest neighbors, we define a ranking position map function as:

$$\text{RP}(T_j^q) = \begin{cases} k, & I_k^j = I_q, I_k^j \in R^{\text{T2I}}(T_j^q, K) \\ K+1, & \text{other} \end{cases} \tag{19}$$

Algorithm 1 Weighted multimodal re-rank for cross-unimodal retrieval.

Input: The similarity matrices S^{I2T}, S^{T2I}, and S^{AT2I}, the size of the testing set M, the number of nearest neighbors K, hyper-parameters w_1 and w_2.

1: **for** $q = 1, 2, ..., M$ **do**
2: Select K-nearest neighbor texts $R^{\text{I2T}}(I_q, K)$ according to the qth row $\boldsymbol{s}_q^{\text{I2T}}$ of S^{I2T}.
3: **for** each $T_j^q \in R^{\text{I2T}}(I_q, K)$ **do**
4: Select K-nearest neighbor images $R^{\text{T2I}}(T_j^q, K)$ according to the jth row of S^{T2I}.
5: Select K-nearest neighbor images $R^{\text{AT2I}}(T_j^q, K)$ according to the jth row of S^{AT2I}.
6: **end for**
7: Calculate the ranking position set $p^{\text{T2I}}(I_q)$.
8: Calculate the ranking position set $p^{\text{AT2I}}(I_q)$.
9: Calculate the importance vector $imp_1(I_q)$.
10: Calculate the importance vector $imp_2(I_q)$.
11: Calculate the re-ranked similarity vector $\boldsymbol{s}_q^{\text{I2T}*}$ by (22).
12: **end for**
13: **return** $S^{\text{I2T}*}$.

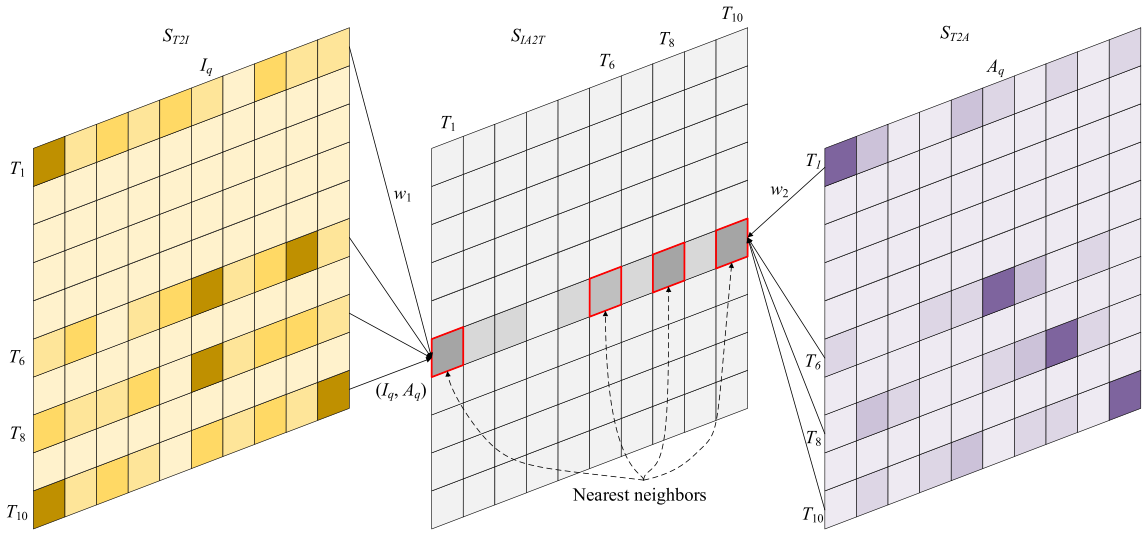


Fig. 4 Proposed weighted multimodal re-rank algorithm

After that, we obtain the ranking position set $p(I_q)$ by applying RP$(\cdot)$ on all testing texts:

$$p(I_q) = \{\mathrm{RP}(T_1^q), \mathrm{RP}(T_2^q), ..., \mathrm{RP}(T_M^q)\} \tag{20}$$

where M is the size of the testing set. According to the definition of RP$(\cdot)$, the smaller RP(T_j^q) is, the more similar T_j^q and I_q are. Therefore, $p(I_q)$ can be considered as a secondary expression of similarity between I_q and each text in the testing set. At this point, we define an importance vector of I_q based on $p(I_q)$ as:

$$imp_1(I_q) = \frac{\{\mathbb{I}(\mathrm{RP}(T_j^q) < K+1)\exp(-\mathrm{RP}(T_j^q))\}_{j=1}^{M}}{\sum_{j=1}^{M}\mathbb{I}(\mathrm{RP}(T_j^q) < K+1)\exp(-\mathrm{RP}(T_j^q))} \tag{21}$$

The importance vector $imp_1(I_q)$ is a modification of $\boldsymbol{s}_q^{\mathrm{I2T}}$ that increases the similarity scores between I_q and those texts which are also similar to I_q. For similarity matrix S^{AT2I} in cross-bimodal retrieval, we can also define the importance vector $imp_2(I_q)$ in the same manner. The final similarity vector $\boldsymbol{s}_q^{\mathrm{I2T}*}$ of I_q is the weighted sum of three vectors:

$$\boldsymbol{s}_q^{\mathrm{I2T}*} = \boldsymbol{s}_q^{\mathrm{I2T}} + w_1 imp_1(I_q) + w_2 imp_2(I_q) \tag{22}$$

We repeat the above procedures for each row in the matrix S^{I2T} to get the WMR similarity matrix $S^{\mathrm{I2T}*}$.

Cross-bimodal Retrieval Here, we take image and audio retrieval text task as an example to illustrate WMR in detail which is shown in Algorithm 2. Given a pair of query image and audio (I_q, A_q) and their corresponding initial similarity vector $\boldsymbol{s}_q^{\mathrm{IA2T}}$ in matrix S^{IA2T}, K-nearest neighbor texts $R^{\mathrm{IA2T}}(I_q, A_q, K) = \{T_1^q, T_2^q, ..., T_K^q\}$ can be collected,

Algorithm 2 Weighted multimodal re-rank for cross-bimodal retrieval.

Input: The similarity matrices S^{IA2T}, S^{T2I}, and S^{T2A}, the size of the testing set M, the number of nearest neighbors K, hyper-parameters w_1 and w_2.

for $q = 1, 2, ..., M$ **do**
  Select K-nearest neighbor texts $R^{\mathrm{IA2T}}(I_q, A_q, K)$ according to the qth row $\boldsymbol{s}_q^{\mathrm{IA2T}}$ of S^{IA2T}.
  for each $T_j^q \in R^{\mathrm{IA2T}}(I_q, A_q, K)$ **do**
    Select K-nearest neighbor images $R^{\mathrm{T2I}}(T_j^q, K)$ according to the jth row of S^{T2I}.
    Select K-nearest neighbor audios $R^{\mathrm{T2A}}(T_j^q, K)$ according to the jth row of S^{T2A}.
  end for
  Calculate the ranking position set $p^{\mathrm{T2I}}(I_q)$.
  Calculate the ranking position set $p^{\mathrm{T2A}}(A_q)$.
  Calculate the importance vector $imp_1(I_q)$.
  Calculate the importance vector $imp_2(A_q)$.
  Calculate the re-ranked similarity vector $\boldsymbol{s}_q^{\mathrm{IA2T}*}$ by (26).
end for
return $S^{\mathrm{IA2T}*}$.

where K is the number of the nearest neighbors. Likewise, we use two similarity matrices, S^{T2I} and S^{T2A}, to perform the reverse search. First for S^{T2I}, the K-nearest images of T_j^q is $R^{\mathrm{T2I}}(T_j^q, K) = \{I_1^j, I_2^j, ..., I_K^j\}$. The ranking position map function for each text is defined as:

$$\mathrm{RP}(T_j^q) = \begin{cases} k, & I_k^j = I_q, I_k^j \in R^{\mathrm{T2I}}(T_j^q, K) \\ K+1, & \text{other} \end{cases} \tag{23}$$

We apply RP$(\cdot)$ on all testing texts to obtain the ranking position set of I_q:

$$p(I_q) = \{\mathrm{RP}(T_1^q), \mathrm{RP}(T_2^q), ..., \mathrm{RP}(T_M^q)\} \tag{24}$$

Based on $p(I_q)$,the importance vector of I_q is defined as:

$$imp_1(I_q) = \frac{\{\mathbb{I}(\mathrm{RP}(T_j^q) < K+1)\exp(-\mathrm{RP}(T_j^q))\}_{j=1}^{M}}{\sum_{j=1}^{M}\mathbb{I}(\mathrm{RP}(T_j^q) < K+1)\exp(-\mathrm{RP}(T_j^q))} \tag{25}$$

Second for S^{T2A}, we also define the importance vector $imp_2(A_q)$ in the same manner. Finally, the similarity vector $\boldsymbol{s}_q^{\mathrm{IA2T}*}$ is the weighted sum of three vectors:

$$\boldsymbol{s}_q^{\mathrm{IA2T}*} = \boldsymbol{s}_q^{\mathrm{IA2T}} + w_1 imp_1(I_q) + w_2 imp_2(A_q) \tag{26}$$

The WMR similarity matrix $S^{\mathrm{IA2T}*}$ is obtained by performing the above procedures on each row of S^{IA2T}.

4 Experiments

4.1 Experimental configuration

4.1.1 Evaluation metric and datasets

The retrieval performance is measured by the widely-used metric named recall at top K (R@K). Three recalls R@1, R@5, and R@10, are reported for all tasks.

The modified Flickr30k [40], COCO [30], and ADE20k [67] datasets we use are provided by [41]. These datasets combine localized narratives (LN) and synchronized speech with the original datasets. Specifically, Flickr30k-LN, COCO-LN, and ADE20k-LN contain the same images as the original versions, but provide a completely different text description for each image. The number of text descriptions for each image is one, whereas the original datasets had five text descriptions per image. Additionally, the datasets with localized narratives have different splits for training and testing sets, as displayed in Table 1. In these datasets, the audio for each image is the pronunciation of the corresponding text description. These text descriptions convey more fine-grained objects and semantics of the images and contain some redundant words. The audios also contain multiple mute clips. These factors increase the difficulty of cross-modal retrieval and are among the reasons for the substantial performance degradation of most models on the three datasets in Section 4.2. We use only the text-image-audio triplets of each dataset.

Table 1 Statistics of three datasets and their mean duration

Dataset	Train Split		Test Split	
	Size	Mean Duration (s)	Size	Mean Duration (s)
Flickr30k	30546	25.84	1023	25.02
COCO	134272	21.95	8573	21.96
ADE20k	20476	20.20	2053	21.27

Due to hardware limitations, we only use the first 14 seconds of each audio in the dataset. On this condition, only those audios whose duration is shorter than 14 seconds can be input in their entirety; otherwise, the exceeding parts will be truncated. Table 1 shows the size and mean audio duration of each dataset (we pre-process the audio data, i.e., remove as many of the mute clips of the audio as possible to increase our available input without adversely affecting the audio content). Figure 5 shows the statistical information about the duration of the audio data. As shown in this figure, merely about 30% of the audio samples in the Flickr30k and COCO datasets can be processed in their entirety, and this proportion is less than 40% in the ADE20k dataset. Nevertheless, this available duration of audio input is long enough to enable our proposed TIAR to perform competitively in retrieval tasks with audio involved. The audio-related retrieval capability of this model can be improved even more if a longer available duration can be used.

4.1.2 Evaluation baseline models

We adopt six state-of-the-art text-image retrieval models as baselines for comparison: ALBEF [28], GSMN [31], BLIP [29], VinVL [64], TCL [57], and X-VLM [60]. We also adopt ACT [39] for text-audio retrieval and TNN-C-CCA [59] for image-audio retrieval. Considering that the audio in each dataset is the pronunciation of its corresponding text, we adopt HUBERT [18], a state-of-the-art speech recognition model, as the baseline for audio-to-text retrieval. Concretely, we input the entire audio data into HUBERT and use the word error rate, a common metric in speech recognition, to evaluate its performance.

For the ease of experiments, we initialize the parameters of our proposed model except the audio encoder using the pre-trained model provided in [60]. For the audio encoder, we use the pre-trained model provided in [10].

4.2 Cross-unimodal retrieval

Tables 2, 3, and 4 present the comparison results of TIAR, TIAR-WMR, and other state-of-the-art methods. These tables reveal that TIAR-WMR achieves state-of-the-art performance in traditional text-image retrieval across all three datasets. Additionally, TIAR achieves competitive retrieval accuracy and outperforms most of the compared models, except for X-VLM, in text-image retrieval. Notably, TIAR still demonstrates promising performance and outperforms the baseline models significantly in retrieval tasks involving audio, particularly in audio-text retrieval. Although TIAR's mean recall in text-image retrieval is around 1% lower than

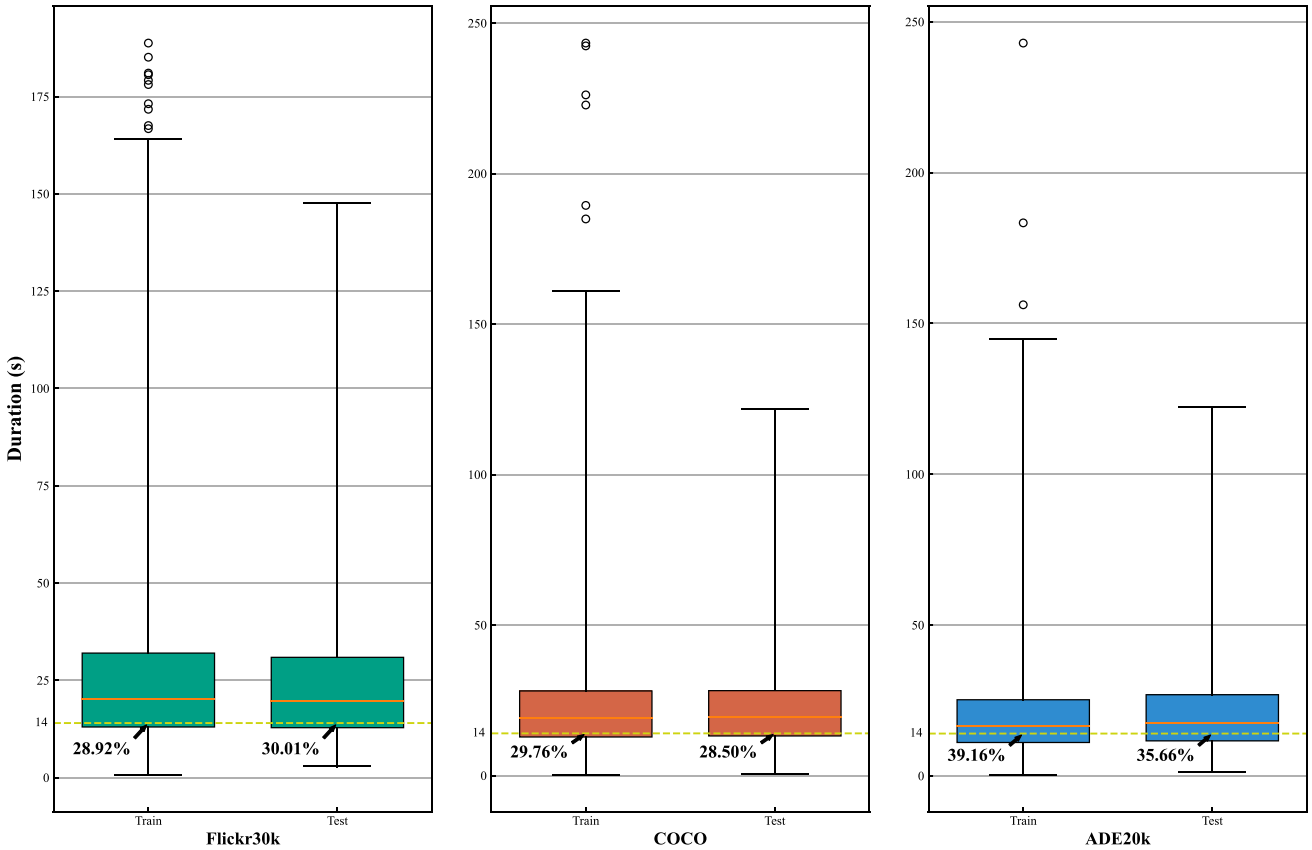


Fig. 5 Box plot of the audio duration statistics for three datasets. The orange line inside each box represents the median duration of the audio. The yellow line marks the position of 14 seconds, and the proportion of audio samples with duration no longer than 14 seconds is annotated

X-VLM, this slight decrease serves as a tradeoff for TIAR's impressive performance in the two audio-involved retrieval tasks. One probable explanation for this decline is that the additional audio-related losses introduced into TIAR act as regularization terms. The most surprising aspect of the results is that both TIAR and TIAR-WMR use only the first 14 seconds of each audio input to achieve their performance, meaning that more than half of the input audios are not processed in their entireties.

4.3 Cross-bimodal retrieval

Table 5 presents the experimental data on cross-bimodal retrieval of TIAR and TIAR-WMR. This table is pretty revealing in three ways. First, with another modality input, all TIAR's retrieval recalls are boosted by varying percentages. In particular, for the image and audio retrieval text task, TIAR significantly outperforms R@1 of the best compared model in the image-to-text task on the Flickr30k, COCO, and ADE20k datasets by 99.51%, 96.98%, and 94.45% respectively, with a significant gap of 47.99% and 29.18% on the latter two datasets. As shown in Fig. 6, the R@1 gain of TIAR in text retrieval with audio input is greater than that of TIAR with just integrating WMR. Though these comparisons are somewhat unfair, they demonstrate TIAR's efficient processing of limited audio and sufficient multimodal fusion with image and audio. Second, TIAR further improves audio retrieval performance with the simultaneous use of text and image. The performance of TIAR in image-audio retrieval is considered relatively poor because of limited audio. However, this performance is improved when both text and image are used simultaneously. The audio retrieval performance of TIAR surpasses that of TIAR with unimodal input remarkably. Third, when WMR is deployed, the performance of TIAR-WMR is still improved in all cross-bimodal tasks and becomes the optimal result on the three datasets, further verifying the superiority of the proposed weighted multimodal re-ranking algorithm.

4.4 Hyper-parameter analysis

We perform three hyper-parameter experiments: parameters of the weighted multimodal re-ranking algorithm, the number of layers in the cross-modal encoder, and the modal combination weights in the cross-bimodal retrieval. All experiments

Table 2 Cross-unimodal retrieval comparison results to TIAR-WMR on Flickr30k dataset. TIAR-WMR achieves state-of-the-art retrieval performance in traditional text-image retrieval

Methods	Time (ms)	Image→Text			Text→Image		
		R@1	R@5	R@10	R@1	R@5	R@10
ALBEF [28]	∼510	66.18	87.98	92.31	65.40	87.29	91.98
GSMN [31]	∼690	73.12	93.03	96.65	74.76	93.54	96.82
BLIP [29]	∼1130	72.82	93.16	96.68	73.12	92.77	96.09
VinVL-base [64]	∼3500	87.53	97.92	98.96	88.77	97.82	98.75
TCL [57]	∼520	62.76	85.14	91.59	61.19	85.53	90.62
X-VLM [60]	∼450	92.96	**99.71**	**99.90**	92.86	99.32	**99.61**
TIAR	∼1140	90.32	99.51	99.71	91.10	98.92	99.51
TIAR-WMR	∼1150	**93.65**	99.51	99.71	**97.17**	**99.41**	99.51
		Text→Audio			Audio→Text		
HUBERT[a] [18]	∼110	-	-	-	24.48	76.11	97.52
ACT [39]	∼700	14.88	37.51	63.58	16.47	37.36	68.26
TIAR	∼1140	76.44	99.22	**99.61**	79.57	99.12	**99.80**
TIAR-WMR	∼1150	**79.77**	**99.61**	**99.61**	**83.94**	**99.61**	**99.80**
		Image→Audio			Audio→Image		
TNN-C-CCA [59]	∼50	5.38	14.77	23.15	5.84	15.31	24.69
TIAR	∼1140	18.57	45.75	61.97	17.11	44.48	61.97
TIAR-WMR	∼1150	**25.40**	**47.89**	**63.28**	**25.79**	**46.85**	**63.95**

[a]word error rate (WER) is less than or equal to 0.2, 0.4, and 0.8 for R@1, R@5, and R@10 hits

are conducted on the Flickr30k dataset and keep the same configuration as the above experiments.

4.4.1 Weighted multimodal re-rank

To analyze the sensitivity of the parameters K, w_1, and w_2 of WMR, we experiment with them. Figure 7 shows the R@1 of TIAR and TIAR-WMR in different tasks (TIAR-WMR for cross-unimodal retrieval and TIAR-WMR+audio for cross-bimodal retrieval). When any of the three parameters increases and the remaining parameters are fixed in cross-unimodal retrieval, R@1 increases slightly. As for cross-bimodal, the R@1 is relatively stable when the three parameters vary.

Table 3 Cross-unimodal retrieval comparison results to TIAR-WMR on COCO dataset. TIAR-WMR achieves state-of-the-art retrieval performance in traditional text-image retrieval

Methods	Time (ms)	Image→Text			Text→Image		
		R@1	R@5	R@10	R@1	R@5	R@10
ALBEF [28]	∼510	34.33	72.29	82.46	34.75	73.29	82.14
GSMN [31]	∼690	39.46	68.00	78.84	42.88	72.22	81.90
BLIP [29]	∼1130	40.90	81.51	90.28	40.45	82.96	90.11
TCL [57]	∼520	33.76	70.57	80.93	32.92	71.42	81.22
X-VLM [60]	∼450	48.99	**90.55**	**96.22**	49.70	91.96	95.93
TIAR	∼1140	47.15	89.09	95.19	47.39	89.51	95.01
TIAR-WMR	∼1150	**49.76**	90.26	95.42	**54.83**	**95.84**	**96.37**
		Text→Audio			Audio→Text		
HUBERT[a] [18]	∼110	-	-	-	20.82	63.44	95.95
ACT [39]	∼700	18.35	41.85	72.40	19.46	42.11	73.84
TIAR	∼1140	90.49	98.40	98.69	92.23	97.23	98.26
TIAR-WMR	∼1150	**96.34**	**98.59**	**98.71**	**95.89**	**98.15**	**98.32**
		Image→Audio			Audio→Image		
TNN-C-CCA [59]	∼50	3.52	12.88	21.45	3.77	13.42	23.79
TIAR	∼1140	15.57	43.81	59.07	15.19	44.77	60.52
TIAR-WMR	∼1150	**24.12**	**48.00**	**62.88**	**24.32**	**48.70**	**63.09**

[a]word error rate (WER) is less than or equal to 0.2, 0.4, and 0.8 for R@1, R@5, and R@10 hits

Table 4 Cross-unimodal retrieval comparison results to TIAR-WMR on ADE20k dataset. TIAR-WMR achieves state-of-the-art retrieval performance in traditional text-image retrieval

Methods	Time (ms)	Image→Text			Text→Image		
		R@1	R@5	R@10	R@1	R@5	R@10
ALBEF [28]	∼510	43.50	70.82	79.69	44.47	71.51	80.18
BLIP [29]	∼1130	47.54	75.35	84.61	51.53	76.18	85.29
TCL [57]	∼520	40.62	68.39	78.18	42.52	69.65	78.57
X-VLM [60]	∼450	65.27	88.55	**93.91**	66.78	89.67	94.06
TIAR	∼1140	63.22	88.07	93.03	65.08	88.75	93.77
TIAR-WMR	∼1150	**69.46**	**88.85**	93.18	**72.58**	**94.11**	**95.23**
		Text→Audio			Audio→Text		
HUBERT[a] [18]	∼110	-	-	-	21.09	59.43	94.69
ACT [39]	∼700	13.66	30.83	64.49	14.63	30.73	62.68
TIAR	∼1140	71.02	95.71	98.00	75.84	94.64	96.69
TIAR-WMR	∼1150	**79.72**	**97.71**	**98.30**	**82.24**	**95.66**	**96.83**
		Image→Audio			Audio→Image		
TNN-C-CCA [59]	∼50	1.92	6.51	10.46	2.01	6.86	11.04
TIAR	∼1140	5.70	22.50	35.41	5.21	22.31	34.83
TIAR-WMR	∼1150	**12.80**	**23.40**	**37.76**	**12.38**	**24.18**	**36.52**

[a]word error rate (WER) is less than or equal to 0.2, 0.4, and 0.8 for R@1, R@5, and R@10 hits

Table 5 Cross-bimodal retrieval results of TIAR-WMR. When inputting two modalities, the retrieval performance of TIAR-WMR is significantly improved

Methods	Flickr30k			COCO			ADE20k		
	Image+Audio→Text			Image+Audio→Text			Image+Audio→Text		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
TIAR	99.51	**99.90**	**99.90**	96.98	98.60	**98.68**	94.45	**97.86**	97.91
TIAR-WMR	**99.61**	**99.90**	**99.90**	**97.49**	**98.61**	98.67	**95.37**	97.71	**97.95**
	Text+Audio→Image			Text+Audio→Image			Text+Audio→Image		
TIAR	91.59	98.92	**99.41**	47.37	89.95	95.12	65.27	88.80	93.86
TIAR-WMR	**93.35**	**99.12**	**99.41**	**49.03**	**90.27**	**95.16**	**68.92**	**89.97**	**94.01**
	Text+Image→Audio			Text+Image→Audio			Text+Image→Audio		
TIAR	92.08	**99.61**	**99.61**	93.47	**98.13**	**98.44**	79.64	96.10	**97.47**
TIAR-WMR	**93.74**	**99.61**	**99.61**	**95.72**	98.06	98.18	**85.05**	**96.49**	97.37

Table 6 Mean recalls of different number of the transformer layers of the cross-modal encoder

Number of layers	Image→Text	Text→Image	Text→Audio	Audio→Text	Image→Audio	Audio→Image
6	95.73	95.67	90.13	90.03	**45.26**	**45.26**
7	**96.51**	**96.51**	**91.76**	**92.83**	42.10	41.19
8	96.35	95.63	86.45	88.56	43.04	42.49

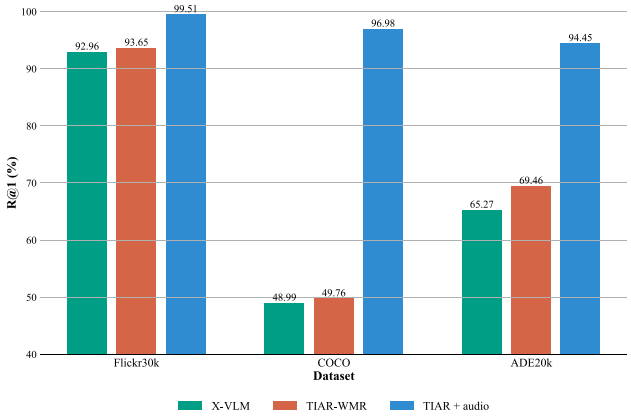


Fig. 6 R@1 of text retrieval of X-VLM and two variants of TIAR on Flickr30k, COCO, and ADE20k datasets

Table 7 α, β, and γ columns report the R@1 and mean recall of image and audio retrieval text, text and image retrieval audio, and text and audio retrieval image in cross-bimodal retrieval, respectively. The mean recall is the mean of R@1, R@5, and R@10

Values	α		β		γ	
	R@1	mR	R@1	mR	R@1	mR
0.1	95.80	98.44	59.24	75.33	59.24	75.37
0.2	98.44	99.32	78.20	88.63	76.25	87.55
0.3	99.02	99.61	86.90	93.91	82.80	92.67
0.4	99.32	99.71	89.64	95.80	88.07	95.05
0.5	**99.51**	**99.77**	91.01	96.71	89.44	95.86
0.6	**99.51**	**99.77**	91.69	96.97	90.32	96.22
0.7	99.12	99.61	**92.08**	**97.10**	89.93	96.09
0.8	98.53	99.35	91.69	96.97	90.81	96.38
0.9	96.58	98.60	89.15	96.12	**91.59**	**96.64**

4.4.2 Number of layers of cross-modal encoder

As stated above, we divide the 12 layers of the BERT-base model into two parts: the first six layers as the text encoder of TIAR and the last six layers as the cross-modal encoder of TIAR. Because of the introduction of audio data, we consider increasing the number of transformer layers of the cross-modal encoder to fuse the three modalities better. Also, it needs to be noted that adding too many transformer layers will significantly increase the complexity of the model and even lead to performance degradation in the case of limited data. Therefore, the number of layers should not be too large.

Table 6 shows the mean recalls of different numbers of the layers of the cross-modal encoder. Our primary concern is the number of retrieval tasks with the highest mean recall for the model. When the cross-modal encoder consists of 7 transformer layers, TIAR achieves the highest mean recall in 4 out of 6 retrieval tasks. Therefore, we set the number of layers of the cross-modal encoder to 7 to achieve the best overall performance.

4.4.3 Modal combination weights

Table 7 shows the R@1 and mean recall of TIAR in cross-bimodal retrieval when the three hyper-parameters in Eq. 9 are set to 9 different values, respectively. The table illustrates that different modal combination weights can lead to drastically different cross-bimodal retrieval results. TIAR obtains the best performance when α is set to 0.5 or 0.6, β is set to 0.7, and γ is set to 0.9. TIAR balanced the similarity information of the two input modalities for α and β. However, for γ, TIAR prefers text-to-image similarity to obtain a better performance due to the low accuracy of audio-to-image retrieval.

4.5 Ablation study

To verify the effectiveness of WMR and the multimodal fusion of TIAR, we conduct an ablation study on the

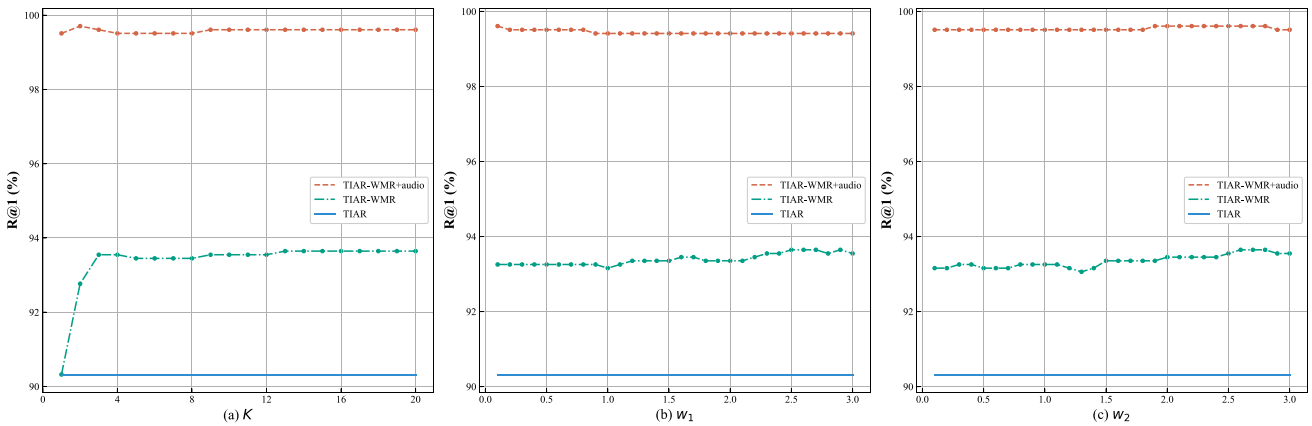


Fig. 7 Sensitivity analysis of K, w_1, and w_2 in WMR

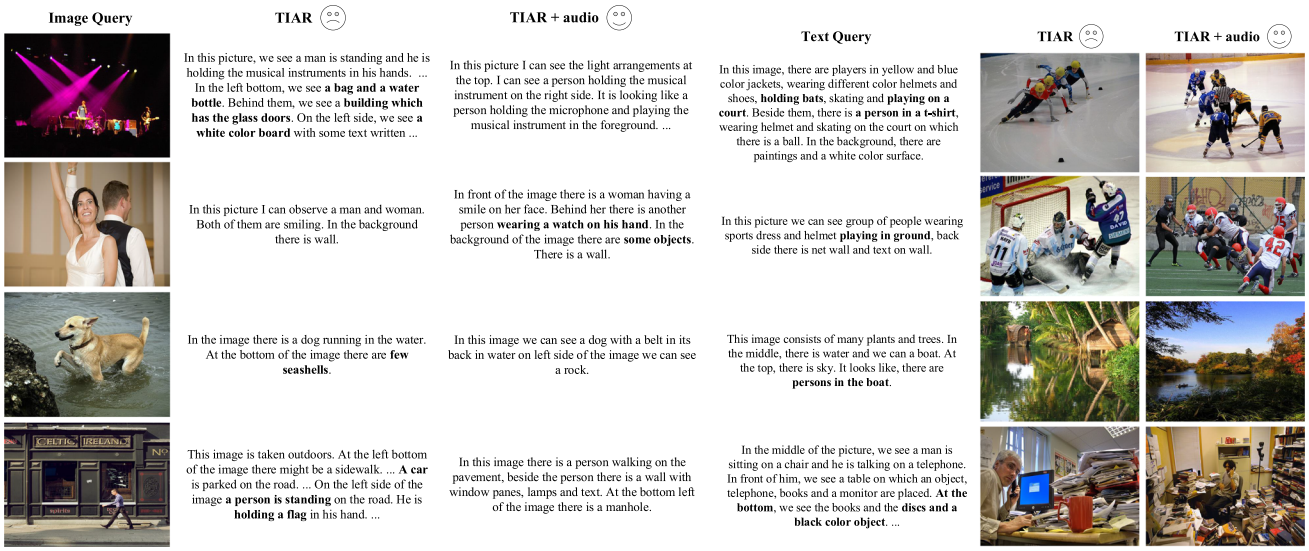


Fig. 8 Qualitative analysis of text and image retrieval for comparisons between the top 1 retrieved results of TIAR with and without audio. Critical text descriptions are bold. Better viewed with zoom-in

Flickr30k dataset. Specifically, we ablate WMR and one of the three input modalities.

Table 8 shows the experimental data of all variants of TIAR. This table shows that when any of the input modalities is ablated, the retrieval performance drops. For example, TIAR is trained as a traditional text-image retrieval model in the case of audio ablation. The R@1 of TIAR w/o audio in text retrieval is 88.17%, which is lower than that of the original TIAR. When any modality is introduced, the performance improves significantly, strongly demonstrating the capability of multimodal fusion of TIAR. The results also prove the effectiveness of WMR by a performance improvement when WMR is combined with TIAR.

Table 8 Ablation study of TIAR on Flickr30k. Models w/o text, image, and audio are the variants where one input modality and WMR are ablated. Model w/o WMR is the variant where WMR is ablated

Methods	Image→Text			Text→Image		
	R@1	R@5	R@10	R@1	R@5	R@10
w/o audio	88.17	98.92	**99.71**	87.10	98.92	**99.61**
w/o WMR	90.32	**99.51**	**99.71**	91.10	98.92	99.51
TIAR-WMR	**93.65**	**99.51**	**99.71**	**97.17**	**99.41**	99.51
	Text→Audio			Audio→Text		
w/o image	41.15	84.07	92.38	47.12	83.97	91.69
w/o WMR	76.44	99.22	**99.61**	79.57	99.12	**99.80**
TIAR-WMR	**79.77**	**99.61**	**99.61**	**83.94**	**99.61**	**99.80**
	Image→Audio			Audio→Image		
w/o text	11.63	31.67	40.08	9.48	27.08	37.63
w/o WMR	18.57	45.75	61.97	17.11	44.48	61.97
TIAR-WMR	**25.40**	**47.89**	**63.28**	**25.79**	**46.85**	**63.95**

4.6 Qualitative analysis

TIAR benefits from multimodal fusion in learning and has boosted performance during testing. To visually analyze the effectiveness of the multimodal fusion of our model, we inspect some examples that are predicted wrongly by TIAR without audio but correctly by TIAR with audio. As shown in Fig. 8, though the retrieved results have similar semantics in some sentences of the caption or parts of the image, they differ in other detailed descriptions or visual objects, which results in the mistakes of TIAR without audio. For example, "few seashells" are the critical clues between the correct and incorrect results in the third row of Fig. 8(a). These clues are ignored by TIAR without audio but are detected by TIAR with audio, which leads to the different predictions of the two variants of TIAR. This result is because the audio plays the role of prompting and emphasizing in both tasks, allowing the model to distinguish the differences between correct and incorrect samples.

5 Conclusion and future work

In this paper, we proposed TIAR, a novel text-image-audio cross-modal retrieval model, and two cross-modal retrieval tasks, named cross-unimodal and cross-bimodal retrieval, to evaluate the performance of TIAR. A weighted multimodal re-ranking algorithm was devised to improve retrieval accu-

racy without additional training. The experimental results show that TIAR-WMR achieves state-of-the-art performance in traditional text-image retrieval and has a promising performance in the two proposed retrieval tasks on Flickr30k, COCO, and ADE20k datasets. The experiments also demonstrate the impressive multimodal fusion capability of TIAR. The retrieval performance of TIAR-WMR is further boosted on the three benchmarks when two input modalities are integrated. In particular, the text retrieval accuracy of TIAR-WMR is significantly improved when provided with additional audio for only 14 seconds (about 30% of the audio samples in the dataset were processed in its entirety).

We still have the following two directions for future research to consider. First, we will continue our work to enable the model to obtain performance gains in audio-involved retrieval tasks without suffering a drop in text-image retrieval. Second, seeking solutions to the problem of only using the first 14 seconds of audio will also be one of our primary efforts in the future.

Acknowledgements This work is supported by National Nature Science Foundation of China (No. 62262006), Technology Innovation and Application Development Key Project of Chongqing (No. cstc2021jscx-gksbX0058), Zhejiang Lab (No. 2021KE0AB01), Open Fund of Key Laboratory of Monitoring, Evaluation and Early Warning of Territorial Spatial Planning Implementation, Ministry of Natural Resources (No. LMEE-KF2021008), Guangxi Key Laboratory of Trusted Software (No. kx202006).

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L (2018) Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6077–6086
2. Baevski A, Zhou Y, Mohamed A, Auli M (2020) wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems 33:12449–12460
3. Brock A, De S, Smith SL, Simonyan K (2021) High-performance large-scale image recognition without normalization. In: International Conference on Machine Learning, PMLR, pp 1059–1071
4. Chen H, Ding G, Liu X, Lin Z, Liu J, Han J (2020a) Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12655–12663
5. Chen L, Ren J, Chen P, Mao X, Zhao Q (2022) Limited text speech synthesis with electroglottograph based on bi-lstm and modified tacotron-2. Applied Intelligence 52(13):15193–15209
6. Chen YC, Li L, Yu L, El Kholy A, Ahmed F, Gan Z, Cheng Y, Liu J (2020b) Uniter: Universal image-text representation learning. In: European conference on computer vision, Springer, pp 104–120
7. Cheng M, Sun Y, Wang L, Zhu X, Yao K, Chen J, Song G, Han J, Liu J, Ding E, et al. (2022) Vista: Vision and scene text aggregation for cross-modal retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 5184–5193
8. Cho J, Lu J, Schwenk D, Hajishirzi H, Kembhavi A (2020) X-LXMERT: Paint, Caption and Answer Questions with Multi-Modal Transformers. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, pp 8785–8805
9. Chung YA, Zhang Y, Han W, Chiu CC, Qin J, Pang R, Wu Y (2021) W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, pp 244–250
10. Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave E, Ott M, Zettlemoyer L, Stoyanov V (2020) Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, pp 8440–8451
11. Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, pp 4171–4186
12. Diao H, Zhang Y, Ma L, Lu H (2021) Similarity reasoning and filtration for image-text matching. Proceedings of the AAAI conference on artificial intelligence 35:1218–1226
13. Dong X, Zhang H, Dong X, Lu X (2021) Iterative graph attention memory network for cross-modal retrieval. Knowledge-Based Systems 226:107138
14. Dou ZY, Xu Y, Gan Z, Wang J, Wang S, Wang L, Zhu C, Zhang P, Yuan L, Peng N, et al. (2022) An empirical study of training end-to-end vision-and-language transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 18166–18176
15. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
16. He K, Chen X, Xie S, Li Y, Dollár P, Girshick R (2022) Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 16000–16009
17. He P, Wang M, Tu D, Wang Z (2023) Dual discriminant adversarial cross-modal retrieval. Applied Intelligence 53(4):4257–4267
18. Hsu WN, Bolte B, Tsai YHH, Lakhotia K, Salakhutdinov R, Mohamed A (2021) Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing 29:3451–3460
19. Huang Z, Zeng Z, Huang Y, Liu B, Fu D, Fu J (2021) Seeing out of the box: End-to-end pre-training for vision-language representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 12976–12985
20. Jia C, Yang Y, Xia Y, Chen YT, Parekh Z, Pham H, Le Q, Sung YH, Li Z, Duerig T (2021) Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning, PMLR, pp 4904–4916
21. Jiang H, Misra I, Rohrbach M, Learned-Miller E, Chen X (2020) In defense of grid features for visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 10267–10276
22. Jin M, Zhang H, Zhu L, Sun J, Liu L (2022) Coarse-to-fine dual-level attention for video-text cross modal retrieval. Knowledge-Based Systems 242:108354

23. Kang P, Lin Z, Yang Z, Fang X, Bronstein AM, Li Q, Liu W (2022) Intra-class low-rank regularization for supervised and semi-supervised cross-modal retrieval. Applied Intelligence 52(1):33–54
24. Kenton JDMWC, Toutanova LK (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, pp 4171–4186
25. Kim S, Kim G, Shin S, Lee S (2021) Two-stage textual knowledge distillation for end-to-end spoken language understanding. ICASSP 2021–2021 IEEE International Conference on Acoustics. Speech and Signal Processing (ICASSP), IEEE, pp 7463–7467
26. Kim W, Son B, Kim I (2021b) Vilt: Vision-and-language transformer without convolution or region supervision. In: Meila M, Zhang T (eds) Proceedings of the 38th International Conference on Machine Learning, PMLR, Proceedings of Machine Learning Research, vol 139, pp 5583–5594,
27. Kong D, Li X, Wang S, Li J, Yin B (2022) Learning visual-and-semantic knowledge embedding for zero-shot image classification. Applied Intelligence pp 1–15
28. Li J, Selvaraju R, Gotmare A, Joty S, Xiong C, Hoi SCH (2021) Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems 34:9694–9705
29. Li J, Li D, Xiong C, Hoi S (2022) Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning, PMLR, pp 12888–12900
30. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. In: European conference on computer vision, Springer, pp 740–755
31. Liu C, Mao Z, Zhang T, Xie H, Wang B, Zhang Y (2020) Graph structured network for image-text matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10921–10930
32. Liu H, Feng Y, Zhou M, Qiang B (2021) Semantic ranking structure preserving for cross-modal retrieval. Applied Intelligence 51:1802–1812
33. Liu Y, Ji S, Fu Q, Zhao J, Zhao Z, Gong M (2022) Latent semantic-enhanced discrete hashing for cross-modal retrieval. Applied Intelligence pp 1–17
34. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021b) Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 10012–10022
35. Luna-Jiménez C, Griol D, Callejas Z, Kleinlein R, Montero JM, Fernández-Martínez F (2021) Multimodal emotion recognition on ravdess dataset using transfer learning. Sensors 21(22):7665
36. Luna-Jiménez C, Kleinlein R, Griol D, Callejas Z, Montero JM, Fernández-Martínez F (2021) A proposal for multimodal emotion recognition using aural transformers and action units on ravdess dataset. Applied Sciences 12(1):327
37. Ma X, Yang X, Gao J, Xu C (2021) The model may fit you: User-generalized cross-modal retrieval. IEEE Transactions on Multimedia 24:2998–3012
38. Malik M, Malik MK, Mehmood K, Makhdoom I (2021) Automatic speech recognition: a survey. Multimedia Tools and Applications 80(6):9411–9457
39. Mei X, Liu X, Huang Q, Plumbley MD, Wang W (2021) Audio captioning transformer. In: Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021, Barcelona, Spain, pp 211–215
40. Plummer BA, Wang L, Cervantes CM, Caicedo JC, Hockenmaier J, Lazebnik S (2015) Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision, pp 2641–2649
41. Pont-Tuset J, Uijlings J, Changpinyo S, Soricut R, Ferrari V (2020) Connecting vision and language with localized narratives. In: European conference on computer vision, Springer, pp 647–664
42. Qi M, Qin J, Yang Y, Wang Y, Luo J (2021) Semantics-aware spatial-temporal binaries for cross-modal video retrieval. IEEE Transactions on Image Processing 30:2989–3004
43. Qian Y, Bianv X, Shi Y, Kanda N, Shen L, Xiao Z, Zeng M (2021) Speech-language pre-training for end-to-end spoken language understanding. ICASSP 2021–2021 IEEE International Conference on Acoustics. Speech and Signal Processing (ICASSP), IEEE, pp 7458–7462
44. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, et al. (2021) Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, PMLR, pp 8748–8763
45. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems 28
46. Roy A, Saffar M, Vaswani A, Grangier D (2021) Efficient content-based sparse attention with routing transformers. Transactions of the Association for Computational Linguistics 9:53–68
47. Schneider S, Baevski A, Collobert R, Auli M (2019) wav2vec: Unsupervised pre-training for speech recognition. Proc Interspeech 2019:3465–3469
48. Seo S, Kwak D, Lee B (2022) Integration of pre-trained networks with continuous token interface for end-to-end spoken language understanding. ICASSP 2022–2022 IEEE International Conference on Acoustics. Speech and Signal Processing (ICASSP), IEEE, pp 7152–7156
49. Stein BE, Meredith MA (1993) The merging of the senses. The MIT press
50. Stein BE, Meredith MA, Huneycutt WS, McDade L (1989) Behavioral indices of multisensory integration: orientation to visual cues is affected by auditory stimuli. Journal of Cognitive Neuroscience 1(1):12–24
51. Tang C, Ma K, Cui B, Ji K, Abraham A (2022) Long text feature extraction network with data augmentation. Applied Intelligence pp 1–16
52. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Advances in neural information processing systems 30
53. Wang L, He K, Feng X, Ma X (2022) Multilayer feature fusion with parallel convolutional block for fine-grained image classification. Applied Intelligence 52(3):2872–2883
54. Wang T, Xu X, Yang Y, Hanjalic A, Shen HT, Song J (2019) Matching images and text with multi-modal tensor fusion and re-ranking. In: Proceedings of the 27th ACM international conference on multimedia, pp 12–20
55. Wu X, Ji S, Wang J, Guo Y (2022) Speech synthesis with face embeddings. Applied Intelligence 52(13):14839–14852
56. Xu Q, Baevski A, Likhomanenko T, Tomasello P, Conneau A, Collobert R, Synnaeve G, Auli M (2021) Self-training and pre-training are complementary for speech recognition. ICASSP 2021–2021 IEEE International Conference on Acoustics. Speech and Signal Processing (ICASSP), IEEE, pp 3030–3034
57. Yang J, Duan J, Tran S, Xu Y, Chanda S, Chen L, Zeng B, Chilimbi T, Huang J (2022) Vision-language pre-training with triple contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 15671–15680
58. You L, Han F, Peng J, Jin H, Claramunt C (2022) Ask-roberta: A pretraining model for aspect-based sentiment classification via sentiment knowledge mining. Knowledge-Based Systems 253:109511
59. Zeng D, Yu Y, Oyama K (2020) Deep triplet neural networks with cluster-cca for audio-visual cross-modal retrieval. ACM

Transactions on Multimedia Computing, Communications, and Applications (TOMM) 16(3):1–23
60. Zeng Y, Zhang X, Li H (2022) Multi-grained vision language pre-training: Aligning texts with visual concepts. In: Chaudhuri K, Jegelka S, Song L, Szepesvari C, Niu G, Sabato S (eds) Proceedings of the 39th International Conference on Machine Learning, PMLR, Proceedings of Machine Learning Research, vol 162, pp 25994–26009,
61. Zhai X, Kolesnikov A, Houlsby N, Beyer L (2022) Scaling vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 12104–12113
62. Zhang F, Xu M, Xu C (2021) Geometry sensitive cross-modal reasoning for composed query based image retrieval. IEEE Transactions on Image Processing 31:1000–1011
63. Zhang L, Wu X (2022) Latent space semantic supervision based on knowledge distillation for cross-modal retrieval. IEEE Transactions on Image Processing 31:7154–7164
64. Zhang P, Li X, Hu X, Yang J, Zhang L, Wang L, Choi Y, Gao J (2021b) Vinvl: Revisiting visual representations in vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 5579–5588
65. Zhang Y, Zhou W, Wang M, Tian Q, Li H (2020) Deep relation embedding for cross-modal retrieval. IEEE Transactions on Image Processing 30:617–627
66. Zhao J, Zhou X, Shi G, Xiao N, Song K, Zhao J, Hao R, Li K (2022) Semantic consistency generative adversarial network for cross-modality domain adaptation in ultrasound thyroid nodule classification. Applied Intelligence pp 1–15
67. Zhou B, Zhao H, Puig X, Xiao T, Fidler S, Barriuso A, Torralba A (2019) Semantic understanding of scenes through the ade20k dataset. International Journal of Computer Vision 127(3):302-321
68. Zhu L, Tian G, Wang B, Wang W, Zhang D, Li C (2021) Multi-attention based semantic deep hashing for cross-modal retrieval. Applied Intelligence 51:5927–5939

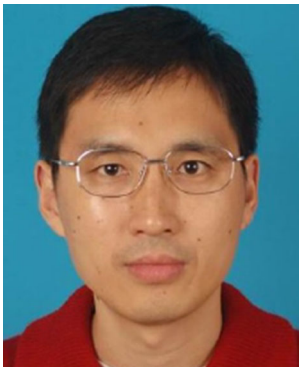
Peide Chi received the B.S. degree in computer science and technology from College of Computer Science at Chongqing University, Chongqing, China, in 2021. Currently he is studying for a master's degree in computer science and technology from Chongqing University, Chongqing, China. His current research interests include Cross-modal retrieval and Multimodal Learning.

Yong Feng received his PhD degree in computer science and technology from College of Computer Science at Chongqing University, Chongqing, China, in 2006. Currently he is a Professor at the College of Computer Science, Chongqing University. His research interest covers Big Data Analysis and Data Mining, Artificial Intelligence and Big Data Processing, Deep Learning and Big Data Retrieval. Corresponding author of this paper.

Mingliang Zhou received the Ph.D. degree in computer science from Beihang University, Beijing, China, in 2017. He was a Postdoctoral Researcher with the Department of Computer Science, City University of Hong Kong, Hong Kong, China, from September 2017 to September 2019. He was a Postdoctoral Fellow with the State Key Lab of Internet of Things for Smart City, University of Macau, Macau, China, from October 2019 to October 2021. He is currently an Associate Professor with the School of Computer Science, Chongqing University, Chongqing, China. His research interests include image and video coding, perceptual image processing, multimedia signal processing, rate control, multimedia communication, machine learning, and optimization.

Xian-cai Xiong received his PhD degree in instrument science and technology at Chongqing University. Currently he is the Vice President and a Senior Engineer at Chongqing Institute of Planning and Natural Resources Investigation and Monitoring. His research interests include spatio-temporal big data and intelligent remote sensing monitoring.

Yong-heng Wang received his PhD degree in computer science and technology from National University of Defense Technology, Changsha, China, in 2006. Currently he is a research specialist at the research center of Big data intelligence, Zhejiang Lab. His research interest covers Big data analysis, machine learning, computer simulation and intelligent decision making.



Bao-hua Qiang received his PhD degree in Department of Computer Science from Chongqing University, Chongqing, China, in 2005. He is now a Professor at the Guangxi Cooperative Innovation Center of cloud computing and Big Data, Guilin University of Electronic Technology. His research interest is Big Data Processing and Information Retrieval.

Authors and Affiliations

Peide Chi[1] · Yong Feng[1] · Mingliang Zhou[1] · Xian-cai Xiong[2,3] · Yong-heng Wang[4] · Bao-hua Qiang[5]

Peide Chi
peide.chi@cqu.edu.cn

Mingliang Zhou
mingliangzhou@cqu.edu.cn

Xian-cai Xiong
xcxiong@126.com

Yong-heng Wang
wangyh@zhejianglab.com

Bao-hua Qiang
qiangbh@guet.edu.cn

1 College of Computer Science, Chongqing University, Chongqing 400030, China

2 Key Laboratory of Monitoring, Evaluation and Early Warning of Territorial Spatial Planning Implementation, Ministry of Natural Resources, Chongqing 400020, China

3 Chongqing Institute of Planning and Natural Resources Investigation and Monitoring, Chongqing 400020, China

4 8# of Zhejiang Lab, Yuhang district, Hangzhou 311121, China

5 Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin 541004, China