

Continual Image-Text Retrieval via Banzhaf Distillation from the Coalitional Game Perspective

Lei Zhao

Beijing University of Technology

Fuhao Wei

Beijing University of Technology

Shaofan Wang

wangshaofan@bjut.edu.cn

Beijing University of Technology

Yanfeng Sun

Beijing University of Technology

Research Article

Keywords: Continual learning, Continual image-text retrieval, Data valuation, Banzhaf distillation, Coalitional game

Posted Date: September 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10760417/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Continual Image-Text Retrieval via Banzhaf Distillation from the Coalitional Game Perspective

Lei Zhao¹, Fuhao Wei¹, Shaofan Wang^{1*}, Yanfeng Sun¹

¹Beijing Key Laboratory of Multimodal Cognitive Computing and Intelligent Software Technology,
School of Artificial Intelligence, Beijing University of Technology, Beijing, China.

*Corresponding author(s). E-mail(s): wangshaofan@bjut.edu.cn;

Contributing authors: zhaolei@emails.bjut.edu.cn; S202273131@emails.bjut.edu.cn;
yfsun@bjut.edu.cn;

Abstract

Continual image-text retrieval (CITR) aims to accomplish the cross-modal retrieval job in the form of sequential multimedia data instead of single entire dataset. Due to the multi-granularity and semantic misalignment of vision and language modalities, CITR exhibits three challenges: spatial disorder of cross-modal data, heterogeneous structure of semantics and hierarchy of knowledge of cross-modal data, and knowledge aliasing during incremental sessions. The key insight is that traditional CITR methods use model the token-by-token similarity, which fails to capture fine-grained semantics especially when knowledge experiences multiple modalities and different incremental sessions. From the coalitional game perspective, we observe that each pairwise tokens of images and texts can act as coalition participants, the Banzhaf valuation can be modeled to characterize multi-grained semantics of image-text pairs. Based on this idea, we propose the **Banzhaf Distillation Network (BDNet)** for CITR. The core component of BDNet is the Banzhaf Distillation module, which preserves the structural cross-modal knowledge with the Banzhaf value, and retains the cross-task structural knowledge with a token-level similarity closeness constraint. Essentially, the Banzhaf valuation models the high order relation of image tokens and text tokens by using their interaction with respect to other tokens towards the semantic alignment of two modalities. Extensive experiments on several single-set scenarios and multi-set scenarios demonstrate the effectiveness of BDNet under single-set and multi-set scenarios. Code is available at <https://github.com/wangsfan/CBD>.

Keywords: Continual learning, Continual image-text retrieval, Data valuation, Banzhaf distillation, Coalitional game

1 Introduction

Multi-modal multimedia data are widely available on the Internet, which greatly affects people’s understanding and perception of the real world. In order to better manage and utilize the cross-modal resources, researchers have made great efforts to link multi-modal data, in order to explore a uniform and robust representation. For instance, with the help of large language models, text description can complement the visual knowledge based on the its non-redundant architecture [1]. Cross-modal retrieval [2–5] is a research topic that retrieves data across different modal types. Among this topic, image-text retrieval which retrieves correctly corresponding image-text pairs, has attracted much attention from scholars in recent years.

In many real-world scenarios, cross-modal data arrive sequentially instead of in one time, bringing new data or even new classes. For instance, doctors conduct medical

diagnosis on patients according to medical record documentation and imaging, which update everyday; they are required to diagnose new and old diseases. Internet produces images and texts everyday, and the network supervision departments are required to find negative contents which appear newly or before. In this sense, the continual image-text retrieval (CITR) task[5, 6], which amounts to retrieve corresponding images (texts) from a text description (an image) in sequential tasks, is significant for real-world applications with limited storage and computational ability.

The CITR involves both cross-domain data learning and continual learning, and hence suffers from three challenges. (a): cross-modal data admit the **multi-granular semantics and spatial disorder**. Without category label prior, the matching clues of images and text descriptions obey different spatial distribution [5], leading to the significant degradation of retrieval performance by

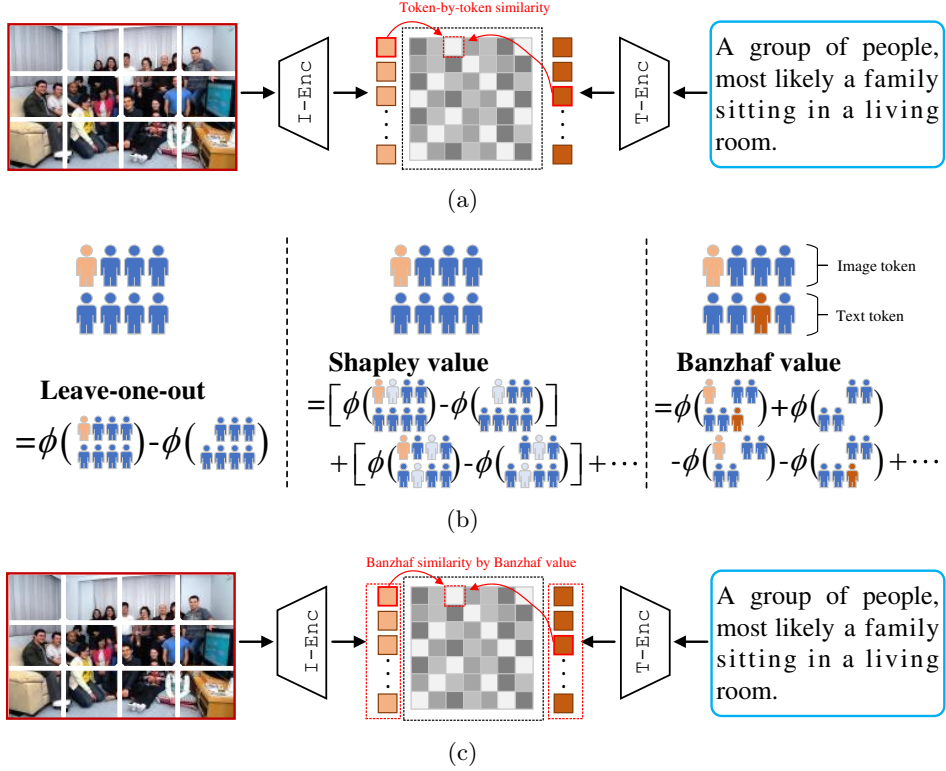


Fig. 1: Motivation of BDNet. (a): traditional CITR methods (e.g., DKR [6] and LPI [10]) model the token similarity matrix with token-by-token similarity, ignoring the effect of the matching token pair on context tokens. (b): our motivation arises from the data valuation work [11] which answers “how much is my data worth”, in the form of three valuation schemes: Leave-one-out error (left), Shapley value (middle), and Banzhaf value (right). (c): we capture fine-grained cross-modal semantics by using intra-task Banzhaf supervision (Section 3.4.2) which involves the context effect by other tokens, and distill fine-grained semantics across different tasks by using inter-task Banzhaf supervision (Section 3.4.3, ignored in the figure).

pre-trained cross-modal models. (b): cross-modal data admit **heterogeneous structure of semantics and hierarchy of knowledge**. The semantically characterized tokens of images differs from those of texts, and the relation among those tokens can be easily jeopardized by improper balancing of old and new knowledge after incremental session. (c): CITR exhibits **knowledge aliasing** during incremental sessions. The matching clues of vision modal and language modal vary during the incremental sessions, including the shift patches of images, words of text descriptions or scale change.

A few researchers attempt to alleviate the aforementioned issues using coarse-to-fine encoders [7], which ignores the evolution of hierarchy of latent spaces of two modals. To circumvent the two challenges simultaneously, we require to learn strong cross-domain aligners to align the multi-granular semantics of different modalities across different task stages during CITR. As a pioneering work on continual cross-modal retrieval, VLKD [8] learns cross-modal knowledge transfer and cross-domain knowledge distillation by using a LSTM-based hierarchical recurrent network. Song and Tan [9] handle the CITR by using a

LSTM-based meta-learner with a task performance constraint. However, the research work on CITR exhibit two shortcomings.

- **Structural semantic ignorance:** LSTM models fine-grained contexts of images and texts by regarding them as sequences, while images and texts exhibit global semantic features which cannot simply modeled by spatial-patch and word-to-word sequences.
- **Inter-task knowledge aliasing:** The knowledge across different multi-modal tasks suffers from serious drifts, in the forms of semantic shift among tokens or scale change of spatial clues.

Recently, transformers [12] achieve great successes in both natural language process and computer vision. Thanks to their tokenized configuration to words and images, transformers can capture high-level structural semantics by learning the self attention over all words and image patches. While several research works [13, 14] utilize transformers for image-text retrieval, the multi-granular semantics of images and texts bring high difficulty to align two modalities by transformers. Even a few models can explore hierarchical knowledge immersed

intermediate layers, a question arise: *how to design tokenized configurations to alleviate the structural semantic ignorance and the inter-task knowledge bias for CITER?*

Recent research work on video-language representation [15] provides a powerful tool: coalitional game theory [16]. Research works on data valuation [11, 17] utilize the coalitional game theory as a tool to answer “how much is my data worth” by regarding each data as a player and modeling the group contribution towards the final task such as image classification. Our key insight is that, each pair of an image token and a text token can act as the role of players for the *token valuation*.

Fig. 1 illustrates the main idea in by using three classic data valuation schemes [11]: leave-one-out error, Shapley value, and Banzhaf value. Among them, the Banzhaf value produces smaller computational cost than the Shapley value, and more generalization power than the leave-one-out error [18]. Essentially, the Banzhaf value characterizes the tendency of each pairwise coalition contributing to the whole data with respect to any subset of the whole data [see detailed definition in (15)]. By linking the Banzhaf value to the vision-language token similarity, it is expected to model the multi-granular semantics and alleviate the inter-task knowledge aliasing during CITER.

Based on the aforementioned discussion, we propose a CITER framework called Banzhaf Distillation Network (BDNet) for preserving fine-grained structural knowledge during cross-task retrieval. The core component of BDNet is the Banzhaf Distillation (CBD) module. The CBD module first regard image tokens and text words as participants in a cooperative game, and preserve structural cross-modal knowledge by using a multivariate cooperative game model. Then it distills cross-task structural knowledge by using the token-level similarity closeness constraint, enabling the memorization of knowledge from the old domain. To facilitate robust memorization of cross-modal cross-task knowledge during CITER, we incorporate BDNet with dynamic semantic prompt learning, and contrastive learning and distillation, where the former adaptively compensates the cross-modal semantic differences, and the latter compensates coarse-grained sample-level knowledge. Both of two ingredients consolidate and memorize cross-modal cross-task knowledge, improving the stability and plasticity performance for CITER. The contribution of BDNet is threefold.

- BDNet aligns the vision modal and language modal by modeling the Banzhaf value of image tokens and text tokens, which is implemented effectively using a Banzhaf Interaction network, from the cooperative game theory perspective.
- BDNet preserves the structural knowledge and alleviates the knowledge aliasing during incremental sessions of CITER, by using the inter-task Banzhaf distillation.
- BDNet achieves state-of-the-art performance over single-set and multi-set continual scenarios for CITER.

The paper is organized as follows. Section 2 reviews previous research works on non-continual image-text retrieval and CITER. Section 3 gives the method of BDNet. Section 4 conducts experiments on single-set scenarios and multi-set scenarios. Section 5 concludes this paper.

2 Related Work

We review previous research work on image-text retrieval and CITER as follows.

2.1 Image-Text Retrieval Methods

2.1.1 Non-deep-learning based methods

Traditional cross modal retrieval methods typically embed data from different modalities through linear projection. Canonical correlation analysis (CCA) is a classic method based on common space learning, which maximizes the correlation between two different modalities in a common space by learning two linear mappings [19]. Rasiwasia *et al.* [20] propose a CCA-based method to improve retrieval accuracy by considering cross-modal associations and semantic abstraction. Costa *et al.* [21] further improve the effectiveness of CAA by adding semantic information. Rasiwasia *et al.* [22] propose the kernel cluster canonical correlation analysis, which alleviates the constraint of data pairing required by CCA. Ranjan *et al.* [23] propose multi-label canonical correlation analysis, which solves the problem of cross modal retrieval with multi-label annotations.

2.1.2 Deep learning methods

Coarse-grained matching methods [24–26] use global features to express images and sentences, embed these features into a common space, and constrain the distance between samples with suitable metric. Frome *et al.* [27] use a linear mapping to align image features and text features in common space. Liu *et al.* [28] propose a recurrent residual fusion block to collect visual and textual representations into a more discriminative embedding space. Wang *et al.* [29] propose a dual branch network, in which the embedded subnetwork aligns the image and text features in the public space, and the similarity subnetwork predicts the modal similarity through the full connection layer. Li *et al.* [30] propose an identity aware two-stage text vision matching framework, which achieves robust image-text retrieval by introducing identity information. Wang *et al.* [31] combine a cross-view ranking constraint with a within-view neighborhood structure preservation constraint, and use two-branch neural network with multi-layer linear projection for image text retrieval. Zheng *et al.* [32] construct an end-to-end dual path convolution network to learn image and text representation. Jian *et al.* [33] propose a deep pairwise ranking model that utilizes the multi-label information to reduce the semantic gap between image features and text descriptions.

Fine-grained matching methods guide visual and text representation learning by focusing on local and structural information [34, 35]. Karpathy *et al.* [36] develop a finer level method to embedded image regions and sentence segments into a common space. MFA [37] proposes a multi-scale fine-grained alignment network to reduce cross modal discrepancy by exploring the multi-scale visual text correspondence. Wang *et al.* [38] propose a dual path structural network based on cross modal prior knowledge, which effectively utilizes complementary modal relationships. HREM [39] utilizes a hierarchical relation model to learn discriminative cross-modal embedding, and segment level and instance level relationships. Diao *et al.* [40] develop two regulators with simple structure to achieve precise fine-grained alignment between modes by learning fine-grained attention. Xiong *et al.* [41] propose an image text matching adaptive network based on reference perception, which solves the problem of cross modal information redundancy. SMAN [42] presents a stacked multimodal attention network with a stacked multimodal attention mechanism to mine the fine-grained relationship between images and texts. UARDA [43] improves semantic alignment by maximally distinguishing relevant and irrelevant word-region fragment features. HBI [44] models a video-text pair as game players and builds the correspondence between text words and video frames with diverse granularity.

2.2 Continual Image-Text Retrieval Methods

Wang *et al.* [45] identify several factors leading to the forgetting in continual cross-modal retrieval, including embedding networks, embedding misalignment, and task overlap. Song and Tan [9] investigate the sequential cross-modal learning task by designing a meta-learner over multiple tasks, with a better performance constraint to absorb the feedback from old and new modalities. Analogue to human cognitive process, KDR [46] decomposes multi-modal features into general and task specific knowledge features, and stores a task specific knowledge extractor for each task to support continuous learning of long sequences. In addition, KDR uses knowledge distillation to replay previous cross modal general knowledge during the learning process, thereby maintaining memory of old knowledge when learning new knowledge. Towards the lifelong histopathological image retrieval task, LWSR [47] designs a distance consistency rehearsal module to anti-forgetting old knowledge, based on the invariant distance between instances of old tasks after learning new tasks. CLCMR [48] designs multi-layer domain selective attention to automatically select parameters specific to previous tasks. In addition, CLCMR utilizes a weight regularization scheme on task specific keys to alleviate excessive drift in sample features and relationships. C²MR [49] proposes a relationship-aware replay strategy to replay the most representative and discriminative samples in the old task feature space, and

effectively reduces catastrophic forgetting by constraining consistency across modal relationships. DKR [6] proposes to filter and rectify affinity knowledge between images and texts, to avoid the negative transfer towards those data which are not retrieved by the old or new model. LPI [10] utilizes multi-modal correlation modules as the components of transformers, and presents a low-rank interaction-augmented decomposition to reduce the parameter size for dynamic expansion of network layers. C-CLIP [50] accomplishes the vision-language continual learning by using contrastive knowledge consolidation with low rank adaptation based on CLIP.

In summary, existing CITR works focus on modeling high order relations of images and texts and maintain knowledge transfer between them, but ignore the subtle token-level modeling of semantics and relation of images and texts, especially from the global semantic correspondence perspective. As a result, when the knowledge exhibits diverse structure during incremental sessions, the shift or scale change of vision and language clues are difficult to characterize, leading to severe forgetting.

3 The Method

We introduce BDNet in this section. Section 3.1 formulates the CITR task and gives an overview of BDNet. Section 3.2 introduces the image encoder and text encoder, serving as the backbone of BDNet. Sections 3.3-3.4 give the details of two modules of BDNet, serving as the key ingredients of BDNet.

We use bold small and bold capital letters to denote vectors and matrices, respectively. We denote $[\cdot; \cdot]$ to be the concatenate operator along column dimension. We denote $\langle \cdot, \cdot \rangle$ to be the cosine similarity. For a matrix $\mathbf{A}$, $[\mathbf{A}]_{i,j}$ denotes its (i,j) th element, and $[\mathbf{A}]_{i,:}$ denotes its i th row vector.

3.1 Preliminaries

3.1.1 Task Formulation

Let $\mathcal{D}^1 \dots \mathcal{D}^K$ be a sequence of image-text pairs with the total number K of task domains, where $\mathcal{D}^k = \{(I_i^k, T_i^k)\}_{i=1}^{|\mathcal{D}^k|}$, with I_i^k denoting the i th image, and T_i^k denoting the text description of the i th image. For $1 \leq k \neq k' \leq K$, $\mathcal{D}^k$ and $\mathcal{D}^{k'}$ obey different domain distributions. For each task $k = 1 \dots K$, by denoting $\mathcal{M}^k \subseteq \cup_{j=1}^k \mathcal{D}^j$ to be the rehearsal data, we learn Θ^k by accessing $\mathcal{D}^k \cup \mathcal{M}^k$ and the previous model Θ^{k-1} . During the inference, we adopt each model Θ_k to retrieve from $\mathcal{D}^{k'}$, $1 \leq k' \leq k \leq K$. CITR aims to learn a sequence of models $\Theta^1 \dots \Theta^K$ towards $\mathcal{D}^1, \dots, \mathcal{D}^K$ respectively, and ultimately enforce the final model Θ^K to retrieve image-text pairs over all the test data of $\cup_{k=1}^K \mathcal{D}^k$.

3.1.2 Overview of BDNet

Fig. 2 gives a framework of BDNet. During each task, an image-text pair is fed into the image encoder and text encoder with semantically enhanced prompts (Section 3.2), the contrastive learning and distillation module (CLD module, Section 3.3), and the Banzhaf distillation module (CBD module, Section 3.4).

3.2 Image Encoder and Text Encoder (IETE) Module

3.2.1 Non-prompt image encoder and text encoder

We first introduce the non-prompt image and text embeddings produced by the image encoder I-Enc and the text encoder T-Enc. In I-Enc, for each image-text pair (I, T) of the k th task, we feed $I \in \mathbb{R}^{H \times W \times 3}$ into a flatten operator $\text{Flat} : \mathbb{R}^{H \times W \times 3} \rightarrow \mathbb{R}^{L_I \times 3p^2}$, a patch embedding layer $\text{PatchEmbed} : \mathbb{R}^{L_I \times 3p^2} \rightarrow \mathbb{R}^{L_I \times d_I}$, the L -layer transformer $\text{Trans}^{(l)} : \mathbb{R}^{(L_I+1) \times d_I} \rightarrow \mathbb{R}^{(L_I+1) \times d_I}$, and a fully connected layer $\text{FC} : \mathbb{R}^{(L_I+1) \times d_I} \rightarrow \mathbb{R}^{(L_I+1) \times d}$ to obtain the *non-prompt image embedding* $\mathbf{I}' \in \mathbb{R}^{(L_I+1) \times d}$, where $p \times p$ denotes the patch size, $L_I = HW/p^2$ denotes the image token number, d_I denotes the image token dimension, and d denotes the output dimension of the image encoder. That is,

$$\mathbf{I} = \text{PatchEmbed} \circ \text{Flat}(I), \quad (1)$$

$$\mathbf{I}' = \text{FC} \circ \text{Trans}^{(L)} \circ \dots \circ \text{Trans}^{(1)}([\mathbf{i}_{\text{cls}}; \mathbf{I}]), \quad (2)$$

where $\mathbf{i}_{\text{cls}} \in \mathbb{R}^{d_I}$ is the learnable class token of images, and the l th transformer layer $\text{Trans}^{(l)}$ utilizes the multi-head self-attention architecture, i.e.,

$$\begin{aligned} \mathbf{X}'' &= \text{Trans}^{(l)}(\mathbf{X}) = \text{MLP}^{(l)}(\text{Norm}(\mathbf{X}')) + \mathbf{X}' \in \mathbb{R}^{(L_I+1) \times d}, \\ \mathbf{X}' &= \text{MSA}^{(l)}(\text{Norm}(\mathbf{X})) + \mathbf{X} \in \mathbb{R}^{(L_I+1) \times d}, \end{aligned} \quad (3)$$

where $\text{MSA}^{(l)}, \text{MLP}^{(l)}, \text{Norm}$ denote the l th multi-head self-attention layer, the l th multilayer perceptron layer, layer normalization, respectively, $l = 1 \dots L$. In T-Enc, we tokenize the text description T by the byte-level Byte-Pair Encoding [51] $\text{BPE} : \{\text{sentences}\} \rightarrow \mathbb{R}^{L_T \times d}$, and then feed the output into the BERT model [52]: $\text{BERT} : \mathbb{R}^{(L_T+1) \times d} \rightarrow \mathbb{R}^{(L_T+1) \times d}$ to obtain the *non-prompt text embedding* $\mathbf{T}' \in \mathbb{R}^{(L_T+1) \times d}$, where L_T denotes the pre-set text token number, and d is the text token dimension which is set as the same as the output dimension of the image encoder. That is,

$$\mathbf{T}' = \text{BERT}([\mathbf{t}_{\text{cls}}; \mathbf{T}]), \quad \text{with } \mathbf{T} = \text{BPE}(T), \quad (4)$$

where $\mathbf{t}_{\text{cls}} \in \mathbb{R}^d$ is the learnable class token of texts.

3.2.2 Prompted image encoder and text encoder

To facilitate fine-grained image-text interaction, we explore inter-modal semantic differences by introducing multiple additional learnable tokens. Those tokens, known as the prompt, can dynamically aggregate image and text information, while retaining the advance of local context matching in the token-wise baseline. Specifically, for the image encoder and text encoder, we add a learnable image prompt $\mathbf{P}_I^k \in \mathbb{R}^{N \times d_I}$ for task k to the output of the patch embedding layer, and a learnable text prompt $\mathbf{P}_T^k \in \mathbb{R}^{N \times d}$ for task k to the output of the Byte-Pair Encoding layer, and feed them into the L -layer transformer $\text{Trans}^{(l)}$, the BERT, respectively, to obtain the *prompted image embedding* $\mathbf{I}''$ and *prompted text embedding* $\mathbf{T}''$. That is,

$$\mathbf{I}_{\text{prompt}} = [\mathbf{P}_I^k; \mathbf{i}_{\text{cls}}; \mathbf{I}] \in \mathbb{R}^{L'_I \times d_I}, \quad \text{with } L'_I := L_I + N + 1, \quad (5)$$

$$\mathbf{T}_{\text{prompt}} = [\mathbf{P}_T^k; \mathbf{t}_{\text{cls}}; \mathbf{T}] \in \mathbb{R}^{L'_T \times d}, \quad \text{with } L'_T := L_T + N + 1, \quad (6)$$

$$\mathbf{I}'' = \text{FC} \circ \text{Trans}^{(L)} \circ \dots \circ \text{Trans}^{(1)}(\mathbf{I}_{\text{prompt}}) \in \mathbb{R}^{L'_I \times d}, \quad (7)$$

$$\mathbf{T}'' = \text{BERT}(\mathbf{T}_{\text{prompt}}) \in \mathbb{R}^{L'_T \times d}. \quad (8)$$

3.2.3 Prompt-key matching strategy for prompt selection

Learning suitable image prompt $\mathbf{P}_I^k$ and text prompt $\mathbf{P}_T^k$ should obey two rules. (a): they can cover rich semantics of each task during training; (b): they can be selected correctly when encountering samples of different tasks during test. Merely using learnable prompts can satisfy the first rule, but not the second one. Alternatively, we learn a task-dependent key corresponding to each prompt to absorb the second rule. In this sense, each task-dependent key, as a low-dimensional vector, serve as an easily-retrieved agent of prompt.

We learn each task-dependent key using a prompt-key matching strategy [53]. Typically, for each batch $\mathcal{B}$ we associate the prompt $\mathbf{P}_I^k$ ($\mathbf{P}_T^k$, resp.) with a d -dimensional learnable key $\mathbf{k}_I^k$ ($\mathbf{k}_T^k$, resp.) by minimizing the similarity with the image (text, resp.) non-prompt class token, i.e.,

$$\mathcal{L}_{\text{prompt}} = \sum_{(I, T) \in \mathcal{B}} (\langle \mathbf{I}'_{\text{cls}}(I), \mathbf{k}_I^k \rangle + \langle \mathbf{T}'_{\text{cls}}(T), \mathbf{k}_T^k \rangle) \quad (9)$$

where $\mathbf{I}'_{\text{cls}}(I) \in \mathbb{R}^d$ is the image non-prompt class token of I [the first row of $\mathbf{I}'$, (2)], and $\mathbf{T}'_{\text{cls}}(T) \in \mathbb{R}^d$ is the text non-prompt class token of T [the first row of $\mathbf{T}'$, (4)].

3.3 Contrastive Learning and Distillation (CLD) Module

During the CITR, we need to guarantee two principles: align the image modal and text modal (intra-task), and preserve old knowledge for alleviating catastrophic forgetting (inter-task). The contrastive learning and distillation (CLD) module accomplishes two jobs based on contrastive learning.

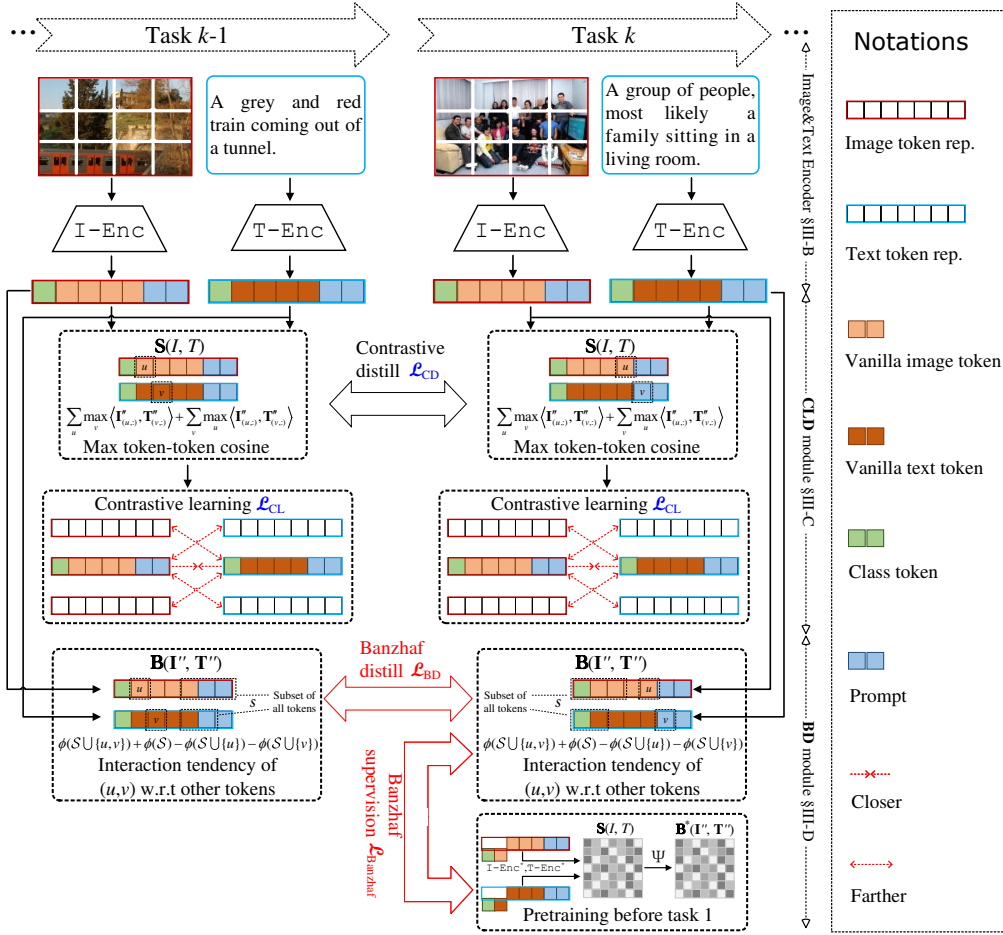


Fig. 2: The overall framework of BDNet. Each image-text pair (I, T) is fed into the image encoder I-Enc and text encoder T-Enc with semantically enhanced prompts (Section 3.2), the contrastive learning and distillation module (CLD module, Section 3.3), and the Banzhaf distillation module (CBD module, Section 3.4).

Specifically, based on the prompted image embedding $\mathbf{I}''$ and prompted text embedding $\mathbf{T}''$, we obtain the *token-wise similarity matrix* $\mathbf{S}(I, T)$ and *similarity quality* $\mathbb{S}(I, T)$ as

$$\mathbf{S}(I, T) = \mathbf{I}''(I) \cdot (\mathbf{T}''(T))^T \in \mathbb{R}^{L'_I \times L'_T}, \quad (10)$$

$$\mathbb{S}(I, T) = \frac{1}{2} \left(\sum_{u=1}^{L'_I} \max_v [\mathbf{S}(I, T)]_{u,v} + \sum_{v=1}^{L'_T} \max_u [\mathbf{S}(I, T)]_{u,v} \right), \quad (11)$$

respectively¹. Then, the image-text contrastive learning loss $\mathcal{L}_{CL}$ is given by

$$\begin{aligned} \mathcal{L}_{CL} &= \mathcal{L}_{i2t} + \mathcal{L}_{t2i}, \\ \mathcal{L}_{i2t} &= -\frac{1}{|\mathcal{B}|} \sum_{(I, T) \in \mathcal{B}} \log \frac{\exp(\mathbb{S}(I, T))}{\sum_{T' \in \mathcal{B}} \exp(\mathbb{S}(I, T'))}, \\ \mathcal{L}_{t2i} &= -\frac{1}{|\mathcal{B}|} \sum_{(I, T) \in \mathcal{B}} \log \frac{\exp(\mathbb{S}(I, T))}{\sum_{I' \in \mathcal{B}} \exp(\mathbb{S}(I', T))}, \end{aligned} \quad (12)$$

where the losses $\mathcal{L}_{i2t}, \mathcal{L}_{t2i}$ contrast the current image-text pair (I, T) towards non-matching texts T' or images I' .

For the second job, as cross-modal knowledge reflects the semantic correlation between image domain and text domain, a natural way to preserve such knowledge is to maintain the invariance of the sample-by-sample image-to-text similarity matrix across different tasks. Thus, let us maintain the closeness of the image-text similarity matrix among rehearsal data $\mathcal{M}^k$ under the $(k-1)$ th and k th tasks, denoted as $\mathbf{S}^{k-1}, \mathbf{S}^k \in \mathbb{R}^{|\mathcal{M}^k| \times |\mathcal{M}^k|}$:

$$[\mathbf{S}^{k-1}]_{i,j} = \mathbb{S}^{k-1}(I_i, T_j), [\mathbf{S}^k]_{i,j} = \mathbb{S}^k(I_i, T_j), I_i, T_j \in \mathcal{M}^k, \quad (13)$$

where $\mathbb{S}^{k-1}(I_i, T_j), \mathbb{S}^k(I_i, T_j)$ represent the similarity quality (11) under previous model Θ^{k-1} and current model Θ^k , respectively. Then, the image-text contrastive distillation loss $\mathcal{L}_{CD}$ is given by

$$\mathcal{L}_{CD} = \text{KL}(\text{softmax}(\mathbf{S}^{k-1}) \| \text{softmax}(\mathbf{S}^k)), \quad (14)$$

where $\text{softmax}(\cdot)$ denotes the softmax operation along the row dimension, and each similarity matrix is vectorized before inputting in the $\text{KL}(\cdot, \cdot)$ function.

¹Essentially, the similarity quality serves as an alternative for the character function ϕ in the Banzhaf value (15) later.

3.4 Banzhaf Distillation (CBD) Module

3.4.1 Revisit Banzhaf valuation

Cooperative game theory [16] consists of a participant set $\mathcal{N} = \{1 \dots N\}$ with characteristic function ϕ , which maps each team of players to a real number indicating the gain that all the players get from completing the task together. In cooperative games, some players tend to form coalitions: there may be situations in which $\phi(u, v)$ is large though $\phi(u)$, $\phi(v)$ are small. The Banzhaf valuation [54] measures the additional benefit from the target coalition and the cost to these participants of losing the coalition with others. Given a coalition $\{u, v\} \subseteq \mathcal{N}$, the Banzhaf value $\mathbb{B}(u, v)$ for two players $\{u, v\}$ with a characteristic function ϕ is defined as:

$$\mathbb{B}(u, v) = \sum_{\mathcal{S} \subseteq \mathcal{N} \setminus \{u, v\}} \frac{1}{2^{N-2}} [\phi(\mathcal{S} \cup \{u, v\}) + \phi(\mathcal{S}) - \phi(\mathcal{S} \cup \{u\}) - \phi(\mathcal{S} \cup \{v\})]. \quad (15)$$

Intuitively, $\mathbb{B}(u, v)$ reflects the tendency of the interactions of $\{u, v\}$ with any subset $\mathcal{S}$ of $\mathcal{N}$. In the image-text retrieval task, a higher value of $\mathbb{B}(u, v)$ indicates that the u th patch of an image and the v th word of a text description approach together.

3.4.2 Intra-task Banzhaf supervision

The CBD module enforces the cross-task cross-modal semantic consistency of image and text tokens in two phases. We first supervise the Banzhaf value evaluation using a pre-trained network during task $\mathcal{D}^k$, and then distill the Banzhaf relation between tasks $\mathcal{D}^{k-1}$ and $\mathcal{D}^k$ for anti-forgetting. The former maintains the model to inherit prior knowledge of Banzhaf values (known as the *intra-task Banzhaf supervision*), while the latter update and filter task-dependent knowledge (known as the *inter-task Banzhaf distillation*).

Typically, for each (I, T) pair of $\mathcal{D}^k$, we model the Banzhaf value $\mathbb{B}(u, v)$ of each token pair (u, v) of (I, T) , and then define $\mathcal{I}_u$ to be the *interaction intensity* of the u th token of I , i.e.,

$$\mathcal{I}_u = \left\{ p_{u,v} = \frac{\exp(\mathbb{B}(u, v))}{\sum_{l=1}^{L'_T} \exp(\mathbb{B}(u, l))} \right\}_{v=1}^{L'_T}, \quad (16)$$

$$\mathbb{B}(u, v) = [\Psi(\mathbf{S}(I, T))]_{u,v},$$

where $\mathbb{B}(u, v)$ is evaluated through a learnable Banzhaf Interaction (BI) network Ψ based on the token-by-token similarity matrix $\mathbf{S}(I, T)$. In essence, $\mathcal{I}_u$ measures the interaction of the u th token of I with all T 's tokens. Then, we supervise $\mathcal{I}_u$ by absorbing pre-trained knowledge from its pre-trained edition $\mathcal{I}_u^*$, whose Banzhaf value is given by a pre-trained BI network Ψ^* (see details in Section 3.4.4),

i.e.,

$$\mathcal{I}_u^* = \left\{ p_{u,v}^* = \frac{\exp([\Psi^*(\mathbf{S}(I, T))]_{u,v})}{\sum_{l=1}^{L'_T} \exp([\Psi^*(\mathbf{S}(I, T))]_{u,l})} \right\}_{v=1}^{L'_T}. \quad (17)$$

Similarly, the *interaction intensity of the v th token of T* and its pre-trained version, denoted as $\mathcal{T}_v, \mathcal{T}_v^*$, are given by

$$\mathcal{T}_v = \left\{ q_{u,v} = \frac{\exp(\mathbb{B}(u, v))}{\sum_{l=1}^{L'_I} \exp(\mathbb{B}(l, v))} \right\}_{u=1}^{L'_I}, \quad (18)$$

$$\mathcal{T}_v^* = \left\{ q_{u,v}^* = \frac{\exp([\Psi^*(\mathbf{S}(I, T))]_{u,v})}{\sum_{l=1}^{L'_I} \exp([\Psi^*(\mathbf{S}(I, T))]_{l,v})} \right\}_{u=1}^{L'_I},$$

respectively. The Banzhaf supervision loss $\mathcal{L}_{\text{Banzhaf}}$ is then given by:

$$\mathcal{L}_{\text{Banzhaf}} = \frac{1}{L'_I L'_T |\mathcal{B}|} \sum_{(I, T) \in \mathcal{B}} \sum_{u=1}^{L'_I} \sum_{v=1}^{L'_T} [\text{KL}(\mathcal{I}_u \| \mathcal{I}_u^*) + \text{KL}(\mathcal{T}_v \| \mathcal{T}_v^*)], \quad (19)$$

which requires the learnable Banzhaf values obey prior knowledge induced from token-wise similarity. In this way, the Banzhaf loss $\mathcal{L}_{\text{Banzhaf}}$ establishes the fine-grained semantic alignment between image patches and text words, and enables efficient and accurate cross-modal retrieval.

3.4.3 Inter-task Banzhaf distillation

To preserve the knowledge learned from the fine-grained semantics across different tasks, we propose Banzhaf-based knowledge distillation. The main idea is that, fine-grained knowledge can be preserved by maintaining the closeness of token-wise similarities of replay data over the old model Θ^{k-1} and new model Θ^k . Specifically, we design the inter-task Banzhaf distillation loss $\mathcal{L}_{\text{BD}}$ based on the rehearsal data $\mathcal{M}^k$ as:

$$\mathcal{L}_{\text{BD}} = \frac{1}{L'_I L'_T |\mathcal{M}^k|} \sum_{(I, T) \in \mathcal{M}^k} \sum_{u=1}^{L'_I} \sum_{v=1}^{L'_T} [\text{KL}(\mathcal{I}_u^{k-1} \| \mathcal{I}_u^k) + \text{KL}(\mathcal{T}_v^{k-1} \| \mathcal{T}_v^k)], \quad (20)$$

where the superscripts $k-1, k$ indicate the interaction intensities obtained from the old model, new model, respectively. Through Banzhaf distillation, the fine-grained knowledge in the old domain and the fine-grained cross-modal relation are preserved and transferred to the new domain.

3.4.4 Design of Banzhaf Interaction network Ψ

We design a network Ψ for evaluating the Banzhaf function $\mathbb{B}(u, v)$ to avoid exponentially expensive computation in (15). The network Ψ utilizes a lightweight architecture, with a convolution layer, a single-head self-attention

layer, and a convolution layer, and takes the similarity matrix $\mathbf{S}(I, T)$ of (I, T) as input and outputs the Banzhaf value. Ψ accomplishes the Banzhaf approximation from two aspects. (a): Ψ mimics $\mathbb{B}(u, v)$ using the interaction of the similarity value $[\mathbf{S}(I, T)]_{u, v}$ with the other similarity values via convolution and self-attention operations. (b): During the pre-training of Ψ , we utilize the similarity quality (11) to estimate the Banzhaf value $\mathbb{B}(u, v)$ (Line 8 of Algorithm 1) and supervise the learning of Ψ .

Typically, during the pre-training, we need prepare the ‘‘ground-truth’’ Banzhaf value $\mathbb{B}(u, v)$. As the Banzhaf value is closely related to the token similarity, we evaluate it via (15) in which we require $\mathcal{S}$ contain at least an image token and at least a text token, and estimate those character function values $\phi(\mathcal{S} \cup \{u, v\}), \phi(\mathcal{S}), \phi(\mathcal{S} \cup \{u\}), \phi(\mathcal{S} \cup \{v\})$ using the similarity quality $\mathbb{S}(I, T)$, with the input $\mathbf{S}(I, T)$ replaced by its submatrix corresponding to input tokens of ϕ . Typically, let $\mathcal{N} = \{u_1, \dots, u_{L_I}\} \cup \{v_1, \dots, v_{L_T}\}$ denote the set of all image tokens and text tokens, and let $\mathcal{S} = \{u_{i_1}, \dots, u_{i_m}\} \cup \{v_{j_1}, \dots, v_{j_n}\}$ be a subset of $\mathcal{N}$. The character function value on $\mathcal{S}$ is given by:

$$\phi(\mathcal{S}) = \frac{1}{2} \left(\sum_{u=1}^{L_I} \max_v [\mathbf{S}|_{\mathcal{S}}]_{u, v} + \sum_{v=1}^{L_T} \max_u [\mathbf{S}|_{\mathcal{S}}]_{u, v} \right) \quad (21)$$

where $\mathbf{S}|_{\mathcal{S}} = [\mathbf{S}(I, T)]_{\{i_1 \dots i_m\}, \{j_1 \dots j_n\}}$ is the submatrix of $\mathbf{S}(I, T)$ induced from $\mathcal{S}$. By using this definition, instead of computing all character function values over each subset $\mathcal{S}$, we sort each row and each column of $\mathbf{S}(I, T)$ in one time, and then take the maximum element based on each subset $\mathcal{S}$, leading to a much smaller complexity $\mathcal{O}(L_I L_T (\log L_I + \log L_T))$.

3.5 Training and Test

Algorithms 1 gives the process of training, which includes the pre-training stage and incremental session. During task k , we train BDNet by optimizing the image encoder (PatchEmbed, $\{\text{Trans}^{(l)}\}, \text{FC}$), text encoder (BERT) and prompts $\mathbf{P}_I^k, \mathbf{P}_T^k$ via Adam based on the following total loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{prompt}} + \mathcal{L}_{\text{CL}} + \mathcal{L}_{\text{CD}} + \alpha \mathcal{L}_{\text{Banzhaf}} + \beta \mathcal{L}_{\text{BD}} \quad (22)$$

where α, β are hyper-parameters.

The test of BDNet involves a prompt-key matching stage. Please refer to Algorithm 2 for details.

3.6 Complexity Analysis

We conduct the complexity analysis of BDNet for each sample (I, T) during each task. During our computation, we adopt the assumption $L \ll L_I \approx L_T \approx L'_I \approx L'_T \ll d_I \approx d$.

(1) Pre-training stage. Each image I passes through I-Enc (L -layer transformer $\text{Trans}^{(l)}, l = 1 \dots L$ and the fully connected layer FC), and each text T passes

through T-Enc, which produces the $\mathcal{O}(L_I^2 d_I L + L_I d_I^2 L + L_I d_I d)$, $\mathcal{O}(L_T d^2 L)$ complexity, respectively. Computing $\mathcal{L}_{\text{CL}}$ requires the $\mathcal{O}(L'_I L'_T d)$ complexity. Besides, computing (I, T) 's Banzhaf matrix $\mathbf{B}(I, T)$ (Line 8 of Algorithm 1) requires $\mathcal{O}(L_I L_T (\log L_I + \log L_T))$ complexity. Then the pre-training stage requires $\mathcal{O}(L_I d_I^2 L + L_T d^2 L)$ in total.

(2) IETE module: similar to the pre-training stage, each image I passes through the L -layer transformer $\text{Trans}^{(l)}, l = 1 \dots L$ and the fully connected layer FC, and each text T passes through the BERT. The IETE module produces $\mathcal{O}(L'_I d_I^2 L + L'_T d^2 L)$.

(3) CLD module: the token-wise similarity (10) produces complexity $\mathcal{O}(L'_I L'_T d + L'_T + L'_I)$, where the second and third terms arise from the maximum computation. Hence, for each pair (I, T) , the image-text contrastive learning loss $\mathcal{L}_{\text{CL}}$ requires the $\mathcal{O}(L'_I L'_T d)$ complexity, and the image-text contrastive distillation loss $\mathcal{L}_{\text{CD}}$ requires the $\mathcal{O}(1)$ complexity per pair. Thus, the CLD module requires $\mathcal{O}(L'_I L'_T d)$.

(4) CBD module: each pair pass through the Banzhaf network Ψ , leading to neglectable complexity since it consists of two convolution layers and a single-head self-attention layer. Then, the two KL operators in $\mathcal{L}_{\text{Banzhaf}}$ $\mathcal{O}(L'_T + L'_I)$ for each u, v . Thus, the Banzhaf loss $\mathcal{L}_{\text{Banzhaf}}$ requires $\mathcal{O}(L'_I L'_T (L'_I + L'_T))$ complexity, and the Banzhaf distillation loss $\mathcal{L}_{\text{BD}}$ requires $\mathcal{O}(L'_I L'_T (L'_I + L'_T))$ complexity. Then the CBD module requires $\mathcal{O}(L'_I L'_T (L'_I + L'_T))$ complexity.

In total, the total complexity of BDNet is $\mathcal{O}(L'_I d_I^2 L + L'_T d^2 L)$, and the IETE module contributes the major complexity towards BDNet.

4 Experiments

We conduct comparative experiments in this section, using two types of CITR scenarios: single-set scenarios and multi-set scenarios.

4.1 Experimental Setup

4.1.1 Dataset

Pascal Sentence [58] is a subset of the *Pascal VOC* dataset, which contains 1000 pairs of graphic data, each image corresponds to a manually labeled five-paragraph text description, and the dataset is divided into 20 categories. In the CITR experiment, we take 700 pairs of samples as the training set and 300 pairs of samples as the test set.

Wikipedia [20] has 2866 image-text pairs, which are divided into 10 categories. In the CITR experiment, we randomly divide the dataset into 2292 pairs of samples as training set and 474 pairs of samples as test set. For each text, we take the first sentence for training and testing.

MS-COCO Caption (MS-COCO) [59] is a widely used image caption dataset. It contains 80K training

Algorithm 1: Training of BDNet

```

input   : Auxiliary dataset  $\mathcal{D}'$ , task sequence  $\bigcup_{k=1}^K \mathcal{D}^t$ ,
output : Parameters of fully connected layer  $\mathbf{FC}$ ,
          task-wise image prompt  $\mathbf{P}_I^k$ , text prompt  $\mathbf{P}_T^k$ , Banzhaf Interaction network  $\Psi$ ,
          MaxEpoch
initialize:  $\mathcal{M}^k \leftarrow \emptyset$ 
// Pre-training stage
1 for  $epoch = 1 \dots \text{MaxEpoch}$  do
2   Sample a mini-batch  $\mathcal{B}$  from  $\mathcal{D}'$ ;
3   Update the non-prompt encoders I-Enc and T-Enc
   by minimizing  $\mathcal{L}_{\text{CL}}$ ;
4   foreach  $(I, T) \in \mathcal{B}$  do
5     Compute token-wise similarity matrix  $\mathbf{S}(I, T)$ 
     via (10);
6     Sort each row and each column of  $\mathbf{S}(I, T)$ 
     decreasingly;
7     Compute  $\phi(\mathcal{S}), \phi(\mathcal{S} \cup \{u\}), \phi(\mathcal{S} \cup \{v\}), \phi(\mathcal{S} \cup \{u, v\})$ 
     via (21),  $\mathcal{S} \subseteq \mathcal{N} \setminus \{u, v\}$ ;
8     Compute  $(I, T)$ 's Banzhaf matrix
      $\mathbf{B}(I, T) = [\mathbb{B}(u, v)]_{L_I \times L_T} \in \mathbb{R}^{L_I \times L_T}$  via (15);
9   end
10  Update BI network  $\Psi$  via minimizing
     $\sum_{(I, T) \in \mathcal{B}} \|\Psi(\mathbf{S}(I, T)) - \mathbf{B}(I, T)\|_F^2$ ;
11 end
12  $\Psi^* \leftarrow \Psi$ ;
// Incremental session
13 for  $\text{taskID}: k = 1, \dots, K$  do
14   for  $epoch = 1 \dots \text{MaxEpoch}$  do
15     Sample a mini-batch  $\mathcal{B}$  from  $\mathcal{D}^t \cup \mathcal{M}^k$ ;
16     Update  $\mathbf{FC}, \mathbf{P}_I^k, \mathbf{P}_T^k, \Psi$  by minimizing
      $\mathcal{L}_{\text{total}}$  (22) via Adam;
17     Update  $\mathcal{M}^k$  for rehearsal data;
18   end
19 end

```

images and 5K testing images, where each image has five captions.

Flickr30K [60] contains 31,783 images from the Flickr website, and each image is annotated by 5 sentences. We use 30K images for training and the rest 1K images for test

IAPR TC-12 [61] consists of 20,000 images with corresponding captions collected around the world. We use 15K images for training and the rest 5K images for test.

ECommerce-T2I [62] is a large-scale E-commerce products retrieval dataset. It contains 90K images for training and 5K images for testing, where each image is annotated with one sentence.

RSICD [63] is a remote sensing image retrieval dataset, which contains 10,921 images of 30 scenes. We use 9,828 images for training and the rest 1,093 images for test.

4.1.2 Task partition

For *Pascal Sentence* and *Wikipedia*, we follow [48] and divide these datasets into tasks to produce different task domains, in order to simulate the scenario of incremental data acquisition. For *MS-COCO*, *Flickr30K*, *IAPR TC-12*, *ECommerce-T2I* and *RSICD*, we follow [6] to

Algorithm 2: Test of BDNet (image-text retrieval)

```

input   :  $I^{\text{test}} \in \mathcal{D}^{\text{test}}$ , prompt-key pairs  $\{(\mathbf{P}_I^k, \mathbf{k}_I^k)\}_{k=1}^K$ , final model  $\Theta^K$ 
output  : Ranking score of  $T_m \in \mathcal{D}^{\text{test}}$ ,  $m = 1 \dots |\mathcal{D}^{\text{test}}|$ 
1 Feed  $I^{\text{test}}$  into non-prompt I-Enc to obtain
  non-prompt class token  $\mathbf{i}'_{\text{cls}}$ ;
2 Select the matching key  $\mathbf{k}_I^{k*}$  by
   $k^* = \arg \max_{k \in \{1 \dots K\}} \langle \mathbf{i}'_{\text{cls}}, \mathbf{k}_I^k \rangle$ ;
3 Feed  $I^{\text{test}}$  into prompted I-Enc with the prompt  $\mathbf{P}_I^{k*}$ 
  to obtain prompted class token  $\mathbf{i}_{\text{cls}}(I^{\text{test}})$ ;
4 for  $m = 1 \dots |\mathcal{D}^{\text{test}}|$  do
5   Feed  $T_m$  into prompted T-Enc with the prompt
    $\mathbf{P}_T^{k*}$  to obtain prompted class token  $\mathbf{t}_{\text{cls}}(T_m)$ ;
6 end
7 Rank  $\{\langle \mathbf{i}_{\text{cls}}(I^{\text{test}}), \mathbf{t}_{\text{cls}}(T_m) \rangle\}_{m=1}^{|\mathcal{D}^{\text{test}}|}$ ;

```

Table 1: Sample-wise complexity of BDNet.

Modules	Submodules	Complexity
Pre-train	Image Enc	$\mathcal{O}(L_I^2 d_I L + L_I d_I^2 L + L_I d_I d)$
	Text Enc	$\mathcal{O}(L_T d^2 L)$
	Banzhaf matrix	$\mathcal{O}(L_I L_T (\log L_I + \log L_T))$
IETE	Image Enc	$\mathcal{O}(L_I^2 d_I L + L_I' d_I^2 L + L_I' d_I d)$
	Text Enc	$\mathcal{O}(L_T' d^2 L)$
CLD	$\mathcal{L}_{\text{CL}}$	$\mathcal{O}(L_I' L_T' d)$
	$\mathcal{L}_{\text{CD}}$	$\mathcal{O}(1)$
CBD	$\mathcal{L}_{\text{Banzhaf}}$	$\mathcal{O}(L_I' L_T' (L_I' + L_T'))$
	$\mathcal{L}_{\text{BD}}$	$\mathcal{O}(L_I' L_T' (L_I' + L_T'))$
Total		$\mathcal{O}(L_I' d_I^2 L + L_T' d^2 L)$

construct a CITR scenario with five sequential datasets which verifies the cross-domain antiforgetting ability of methods. Thus, we set up single-set scenarios: Scenario 1, Scenario 2, and multi-set scenarios: Scenarios 3–6), in which Scenarios 4–6 correspond to longer task sequences. Table 2 shows the details of task division.

4.1.3 Implementation details

In our experiments, we initialize the image, text encoder using ViT-B/32 weights. We use ViT-B/16 [12] for feature extraction of images, and the BERT [52] with 12-layers of transformer blocks for feature extraction of texts. Each multi-head attention module for text and image encoders contains 8 and 12 self-attention modules, respectively. We implement BDNet based on Python 3.8 and Pytorch 1.8.6, using an NVIDIA A100 GPU with 80 GB of memory. We use the Adam optimizer with the batch size 64, learning rate 10^{-5} and the training epoch 30. For *Pascal Sentence* and *Wikipedia*, we reserve five and ten image-text pairs

Table 2: Six CITER scenarios with task division.

Scenario	Dataset	Task division	Buffer size
Scenario 1	<i>Pascal Sentence</i>	Task1: airplane, bicycle, boat, bus, car, motorbike, train	5-5
		Task2: bird, cat, cow, dog, horse, person, sheep	
		Task3: dining table, sofa, potted plant, tv monitor, chair, bottle	
Scenario 2	<i>Wikipedia</i>	Task1: royalty, sport, warfare	10-10
		Task2: literature, media, music	
		Task3: art, biology, geography, history	
Scenario 3	<i>Wikipedia</i>	Task1 (W): royalty, sport, warfare	10-10-10
	<i>Pascal Sentence</i>	Task2 (W): literature, media, music	
		Task3 (W): art, biology, geography, history	
		Task4 (P): airplane, bicycle, boat, bus, car, motorbike, train	
Scenario 4	<i>Wikipedia</i>	Task1 (W): royalty, sport, warfare	10-10-10-5
	<i>Pascal Sentence</i>	Task2 (W): literature, media, music	
		Task3 (W): art, biology, geography, history	
		Task4 (P): airplane, bicycle, boat, bus, car, motorbike, train	
		Task5 (P): bird, cat, cow, dog, horse, person, sheep	
Scenario 5	<i>Wikipedia</i>	Task1 (W): royalty, sport, warfare	10-10-10-5-5
	<i>Pascal Sentence</i>	Task2 (W): literature, media, music	
		Task3 (W): art, biology, geography, history	
		Task4 (P): airplane, bicycle, boat, bus, car, motorbike, train	
		Task5 (P): bird, cat, cow, dog, horse, person, sheep	
		Task6 (P): dining table, sofa, potted plant, tv monitor, chair, bottle	
Scenario 6		Task1: <i>MS-COCO</i>	20-20-20-20
		Task2 : <i>Flickr30K</i>	
		Task3: <i>IAPR TC-12</i>	
		Task4: <i>ECommerce-T2I</i>	
		Task5: <i>RSICD</i>	

Table 3: Hyperparameters and modules of BDNet. For two encoders, the output dimension, transformer layer number and prompt number share the same parameter.

Module	Hyperparameters	Symbols	Scenarios 1-6
I-Enc	Image resolution	(H, W)	(224, 224)
	Patch size	$p \times p$	32×32
	Image token dimension	d_I	768
	Vanilla image token number	$L_I = HW/p^2$	49
	Prompted image token number	$L'_I = L_I + N + 1$	55
	Head number	–	12
	Dimension of output	d	512
	Transformer layer number	L	12
	Prompt number	N	5
T-Enc	Text token dimension	d	512
	Vanilla text token number	L_T	77
	Prompted text token number	$L'_T = L_T + N + 1$	–
	Head number	–	8
CBD	BI network	$\Psi : \mathbb{R}^{L'_I \times L'_T} \rightarrow \mathbb{R}^{L'_I \times L'_T}$	CNN (kernel 3×3 , stride 1, padding 1) + single-head self attention layer
Loss	Weights of total loss	$\cdot, \alpha, \beta$	1, 1, 0.1, 0.3

for each class in the memory buffer, respectively. Table 3 lists all the parameters in this paper.

4.1.4 Evaluation metrics

To validate the effectiveness of BDNet, we use mean average precision (mAP) as an evaluation metric. The mAP metric is a widely used performance evaluation criterion

Table 4: Quantitative results on single set scenarios: Scenario 1 and Scenario 2. **Bold** denotes the best, and underline denotes the second best. * denotes fine-grained setting.

	Method	Task1			Task2			Task3			ACC↑	AIA↑	BWT↑
		I2T	T2I	AVER	I2T	T2I	AVER	I2T	T2I	AVER			
Scenario 1	LWF[55]	0.6300	0.6344	0.6322	<u>0.6976</u>	0.6707	<u>0.6842</u>	<u>0.7554</u>	0.7187	<u>0.7370</u>	<u>0.6844</u>	<u>0.6647</u>	-0.0873
	LWF*[55]	0.6118	0.6371	0.6244	0.6762	<u>0.6718</u>	0.6740	0.7392	0.7024	0.7208	0.6730	0.6438	-0.1087
	EWC[56]	0.6257	0.5620	0.5938	0.7056	0.6482	0.6769	0.6480	0.7004	0.6742	0.6483	0.6109	-0.1102
	EWC*[56]	0.6160	0.6249	0.6204	0.6627	0.6492	0.6560	0.7061	<u>0.7277</u>	0.7169	0.6644	0.6436	-0.1056
	RATT[57]	0.4973	0.4669	0.4821	0.5206	0.4376	0.4791	0.3669	0.3638	0.3654	0.4422	0.4361	-0.2113
	CLCMR[48]	<u>0.6569</u>	<u>0.6490</u>	<u>0.6529</u>	0.6486	0.5940	0.6213	0.4543	0.4389	0.4466	0.5736	–	<u>-0.0828</u>
	BDNet	0.8511	0.7922	0.8216	0.9029	0.8467	0.8748	0.8109	0.8712	0.8410	0.8458	0.8329	-0.0116
Scenario 2	LWF[55]	0.5190	0.5007	0.5099	0.4791	0.4529	0.4660	0.6296	0.6502	0.6399	0.5502	0.5225	<u>-0.1675</u>
	LWF*[55]	0.5049	0.5133	0.5091	0.4593	0.4429	0.4511	0.6187	0.6279	0.6233	0.5278	0.5056	-0.1832
	EWC[56]	0.5206	0.5001	0.5104	0.4793	<u>0.4619</u>	<u>0.4706</u>	0.6308	<u>0.6497</u>	0.6402	0.5504	<u>0.5296</u>	-0.1716
	EWC*[56]	0.5372	0.5050	0.5211	0.4714	0.4594	0.4654	0.6433	0.6049	0.6241	0.5368	0.5073	-0.1792
	RATT[57]	0.5234	0.4610	0.4922	0.4543	0.3923	0.4233	0.6301	0.5515	0.5908	0.5021	0.4895	-0.2155
	CLCMR[48]	0.6128	<u>0.5380</u>	0.5754	<u>0.4939</u>	0.4091	0.4515	0.6795	0.6235	<u>0.6515</u>	<u>0.5595</u>	–	-0.1705
	BDNet	<u>0.5434</u>	0.5467	<u>0.5450</u>	0.5105	0.4964	0.5034	<u>0.6718</u>	0.6324	0.6521	0.5668	0.5304	-0.1120

in cross-modal retrieval research [26], which jointly takes into account both ranking information and precision. For CITR, we denote $\text{mAP}_{i,j}$ to be the top-1 mean average precision (i.e. $\text{mAP}@1$) on the i th task $\mathcal{D}^i$ using the model Θ^j , $i \leq j$, and introduce four commonly used metrics: average accuracy (ACC), average incremental accuracy (AIA), and backward transfer (BWT) [46]. That is,

$$\begin{aligned}
 \text{ACC} &= \frac{1}{K} \sum_{k=1}^K \text{mAP}_{k,K}, \\
 \text{AIA} &= \frac{1}{K} \sum_{k=1}^K \left(\frac{1}{k} \sum_{j=1}^k \text{mAP}_{j,k} \right), \\
 \text{BWT} &= \frac{1}{K-1} \sum_{k=1}^{K-1} (\text{mAP}_{k,K} - \text{mAP}_{k,k}),
 \end{aligned} \tag{23}$$

Essentially, ACC evaluates the average accuracy of all tasks tested by the final model Θ^K , AIA evaluates the average accuracy of all previous tasks tested by each model Θ^k , and BWT indicates how much the model forgets the historical tasks after learning new tasks.

4.1.5 Baselines

In order to validate the effectiveness of the model, we choose the current excellent model as the baseline for comparison. LWF [55] uses knowledge distillation to guide the learning of new task data on the current model with the help of soft labels obtained by using new task data output from the old model. EWC [56] uses the Fisher matrix of network weights to estimate the importance of network parameters and imposes penalties on important parameters during training. RATT [57] is a continual image captioning method based on weight regularization, which uses task-specific attention to focus on different tasks

and retains previously learned knowledge through targeted penalization of network parameters. CLCMR [48] is a continual learning method for image text retrieval based on attention regularization. It acquires knowledge from task-relevant and domain-specific attention levels, designs a memory factor for weight regularization, and uses a memory loss function to constrain the choice of attention. In addition, we design fine-grained retrieval paradigms for LWF [55] and EWC [56], and conduct comparative experiments under the same experimental settings.

4.2 Comparison with State-of-the-arts on Single-set Scenarios

Table 4 shows the quantitative results on a single dataset using BDNet and other competitive methods. By analyzing the results, we find that for both Scenario 1 and Scenario 2, compared to LWF[55] and EWC[56], fine-grained LWF/EWC does not show significant performance advantage, which may imply that simply using fine-grained retrieval does not lead to significant performance improvement. For Scenario 1, BDNet achieves optimal values for all metrics on all three tasks. Specifically, BDNet obtains an average mAP of 0.8216, 0.8748, and 0.8410 on Tasks 1–3, which are significantly higher than the state-of-the-art approaches. Moreover, BDNet achieves optimal values for the ACC, AIA and BWT metrics, outperforming the state-of-the-art methods by 0.16, 0.16 and 0.07, respectively. This indicates that BDNet maintains better performance for both historical and new tasks after all tasks are trained.

Let us check Scenario 2. As shown in Table 4, BDNet achieves optimal values on the average mAP for Tasks 2 and 3 in Scenario 2, respectively, with the second position for Task 1. For the ACC, AIA and BWT, BDNet outperforms the state-of-the-art method by 0.007, 0.001, 0.05, respectively. This indicates that for larger

Table 5: Quantitative results of CITR for Scenario 3 (a), Scenario 4 (b) and Scenario 5 (c) on multiple datasets. **Bold** denotes the best, and underline denotes the second best. * denotes fine-grained setting.

(a) Quantitative results for Scenario 3.

Method	task1			task2			task3			task4			ACC↑	AIA↑	BWT↑
	I2T	T2I	AVER	I2T	T2I	AVER	I2T	T2I	AVER	I2T	T2I	AVER			
LWF[55]	0.4170	0.3898	0.4034	0.4062	<u>0.4216</u>	0.4139	0.5767	0.5338	0.5552	0.7144	0.7151	0.7147	0.5218	0.4892	-0.2078
LWF*[55]	0.4326	0.3853	0.4090	0.4152	0.3912	0.4032	0.5009	0.5233	0.5121	<u>0.7265</u>	0.6957	0.7111	0.5088	0.4627	-0.2135
EWC[56]	0.4210	0.4091	0.4150	0.4116	0.3951	0.4033	0.5563	<u>0.5767</u>	<u>0.5665</u>	0.7236	0.7135	0.7185	0.5258	<u>0.4933</u>	-0.1924
EWC*[56]	0.4349	0.3928	0.4138	0.4010	0.3944	0.3977	0.5353	0.5743	0.5548	0.7263	0.6945	0.7104	0.5191	0.4787	-0.1945
RATT[57]	<u>0.5760</u>	<u>0.5157</u>	<u>0.5458</u>	0.4723	0.4011	0.4367	0.4776	0.4411	0.4594	0.7129	<u>0.7210</u>	<u>0.7169</u>	0.5397	0.4911	-0.1692
CLCMR[48]	0.6674	0.5078	0.5876	0.5119	0.3967	<u>0.4543</u>	0.5907	0.5016	0.5462	0.6930	0.7132	0.7031	<u>0.5728</u>	–	<u>-0.1402</u>
BDNet	0.5341	0.5310	0.5325	<u>0.4966</u>	0.4901	0.4934	<u>0.5767</u>	0.5814	0.5791	0.8409	0.7428	0.7918	0.5992	0.5469	0.1290

(b) Quantitative results for Scenario 4.

Method	task1			task2			task3			task4			task5			ACC↑	AIA↑	BWT↑
	I2T	T2I	AVER	I2T	T2I	AVER	I2T	T2I	AVER	I2T	T2I	AVER	I2T	T2I	AVER			
LWF[55]	0.4460	0.3876	0.4168	0.4092	0.3908	0.4000	0.5156	0.5225	0.5190	<u>0.6908</u>	0.6599	<u>0.6753</u>	<u>0.7383</u>	0.7103	0.7243	0.5450	0.4815	-0.2094
LWF*[55]	0.4018	0.4155	0.4086	0.4005	0.3870	0.3937	0.5039	0.4936	0.4987	0.6790	0.6619	0.6704	0.7087	0.7213	0.7150	0.5372	0.4893	-0.2145
EWC[56]	0.4206	0.4194	0.4200	0.4155	0.3968	0.4061	0.5208	<u>0.5381</u>	<u>0.5294</u>	0.6716	0.6524	0.6620	0.7236	0.7431	<u>0.7335</u>	<u>0.5502</u>	<u>0.4951</u>	-0.1927
EWC*[56]	0.4148	0.3913	0.4030	0.3896	0.3907	0.3901	0.4890	0.5122	0.5006	0.6388	<u>0.7029</u>	0.6708	0.7102	<u>0.7445</u>	0.7273	0.5403	0.4828	-0.1952
RATT[57]	0.5250	0.5089	0.5169	0.4277	0.4193	0.4235	0.4520	0.4778	0.4649	0.4535	0.4940	0.4736	0.6734	0.6714	0.6724	0.4385	0.3735	-0.2216
CLCMR[48]	0.6878	<u>0.5239</u>	0.6058	0.5130	<u>0.4080</u>	<u>0.4605</u>	<u>0.5242</u>	0.4820	0.5031	0.6464	0.6800	0.6632	0.6850	0.6799	0.6824	0.5324	–	-0.1240
BDNet	<u>0.5278</u>	0.5182	<u>0.5230</u>	<u>0.5014</u>	0.5021	0.5017	0.5691	0.5718	0.5704	0.7380	0.7489	0.7434	0.7624	0.7619	0.7621	0.6201	0.5659	<u>-0.1280</u>

(c) Quantitative results for Scenario 5.

Method	task1			task2			task3			task4			task5			task6			ACC↑	AIA↑	BWT↑
	I2T	T2I	AVER	I2T	T2I	AVER	I2T	T2I	AVER	I2T	T2I	AVER	I2T	T2I	AVER	I2T	T2I	AVER			
LWF[55]	0.4306	0.3674	0.3990	0.4052	0.3928	0.3990	0.5062	<u>0.5148</u>	<u>0.5105</u>	0.6620	0.6112	0.6366	<u>0.7422</u>	<u>0.6965</u>	<u>0.7193</u>	0.6734	0.7292	0.7013	0.5684	0.5235	-0.2135
LWF*[55]	0.4180	0.3558	0.3868	0.3874	0.3909	0.3891	0.5079	0.4721	0.4900	0.6703	0.6100	0.6431	0.7174	0.6805	0.6989	0.6945	0.7142	0.7043	0.5537	0.5151	-0.2146
EWC[56]	0.4380	0.3970	0.4175	0.4164	0.3946	0.4055	0.5016	0.5135	0.5075	<u>0.6708</u>	0.6244	<u>0.6476</u>	0.7406	0.6796	0.7101	<u>0.7174</u>	0.7331	<u>0.7252</u>	<u>0.5795</u>	<u>0.5381</u>	-0.1752
EWC*[56]	0.4073	0.3942	0.4007	0.3811	0.3983	0.3897	0.4982	0.5060	0.5021	0.6480	<u>0.6337</u>	0.6408	0.7084	0.6935	0.7009	0.6554	<u>0.7552</u>	0.7053	0.5574	0.5219	-0.1861
RATT[57]	0.5560	0.4839	0.5219	0.4309	0.3880	0.4095	0.4480	0.4464	0.4472	0.4416	0.3785	0.4101	0.4764	0.3923	0.4343	0.4195	0.3968	0.4081	0.4385	0.4033	-0.2216
CLCMR[48]	0.6569	<u>0.5076</u>	0.5823	0.5050	<u>0.4089</u>	<u>0.4570</u>	<u>0.5182</u>	0.4828	0.5005	0.6241	0.5922	0.6081	0.6397	0.5801	0.6099	0.4535	0.4196	0.4365	0.5324	–	<u>-0.1267</u>
BDNet	<u>0.5029</u>	0.5027	<u>0.5028</u>	<u>0.4879</u>	0.4845	0.4862	0.5676	0.5698	0.5687	0.7245	0.7167	0.7210	0.7526	0.7508	0.7517	0.7270	0.7464	0.7367	0.6278	0.5840	-0.1245

Table 6: Quantitative results on multi-set scenario: Scenario 6. **Bold** denotes the best, and underline denotes the second best. **Joint** denotes training all data at any time, serving as an upper bound. **Seq** denotes sequentially training and testing the model without anti-forgetting strategies.

Method	MS-COCO			Flickr30K			IAPR TC-12			ECommerce-T2I			RSICD			Average		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Joint	53.7	78.4	86.4	80.9	95.6	98.1	53.6	82.6	90.5	18.6	43.7	57.0	12.0	31.3	45.6	43.7	66.3	75.5
Seq	39.0	65.5	75.3	69.7	90.1	94.3	47.0	77.4	87.0	14.9	37.1	50.1	12.8	33.5	47.9	36.7	60.7	70.9
EWC[56]	39.9	66.6	76.5	70.0	90.6	94.5	47.3	77.3	87.0	15.0	37.2	50.3	13.3	32.5	47.5	37.1	60.8	71.1
LWF[55]	47.0	72.9	82.1	76.2	93.6	<u>97.1</u>	54.6	83.6	91.2	17.0	40.2	53.8	14.0	35.0	49.7	41.7	<u>65.0</u>	<u>74.8</u>
ERL[64]	48.3	73.5	82.6	77.3	93.6	96.7	51.0	80.5	88.8	<u>18.2</u>	42.1	55.5	14.1	33.8	48.2	41.8	64.7	74.3
AFC[65]	48.3	73.7	82.6	77.7	93.6	96.7	51.2	80.7	89.0	18.3	<u>42.2</u>	55.4	14.5	33.3	47.7	<u>42.0</u>	64.7	74.3
Mod-X[5]	<u>49.7</u>	<u>73.8</u>	82.6	<u>77.8</u>	<u>93.7</u>	<u>97.1</u>	51.1	80.7	89.0	<u>18.2</u>	42.3	55.5	14.5	33.5	47.8	41.9	64.8	74.3
DKR[6]	51.4	76.0	84.4	79.3	94.6	97.4	<u>53.9</u>	<u>83.1</u>	<u>90.8</u>	18.0	42.0	54.8	14.1	<u>34.3</u>	<u>49.2</u>	43.3	66.0	75.3
TBDNet	48.9	73.7	<u>82.7</u>	77.3	93.5	96.4	51.0	82.1	89.9	18.1	41.8	54.7	14.1	33.9	48.3	41.9	<u>65.0</u>	74.4

datasets, BDNet produces small gains compared with state-of-the-art methods.

4.3 Comparison with State-of-the-arts on Multi-set Scenarios

Tables 5 and 6 conduct cross-dataset CTR tasks to reflect the model’s ability to resist forgetting and generalization. Table 5a shows the results of Scenario 3, where the data of the first three tasks are from *Wikipedia* and the data of the fourth task are from *Pascal Sentence*. We find that BDNet exhibits excellent performance across datasets. Specifically, BDNet achieves the best average mAP for the last three tasks, with 0.039, 0.013, 0.075 increment compared to the sub-optimal results, respectively. For the ACC, AIA and BWT, BDNet achieves an improvement of 0.026, 0.053, 0.011 compare to the sub-optimal method respectively. This indicates that in the face of large domain differences, BDNet still has sufficient generalization ability to retain knowledge of the old task while learning the new task.

To verify whether BDNet has superior performance on longer task sequences, we conduct experiments on Scenarios 4–6, with 5, 6, 5 sequential tasks, respectively. According to Table 5b, BDNet achieves the optimal value of the average mAP on the latter four tasks of Scenario 4, with 0.041, 0.041, 0.068 and 0.027 higher compared to the sub-optimal methods, respectively. Meanwhile, BDNet achieves the 0.070 increment on ACC and AIA. From Table 5c, BDNet achieves the optimal value on the average mAP for all tasks of Scenario 5 except the first task. Specifically, it outperforms the second position by 0.029, 0.058, 0.073, 0.032, and 0.012 on the mAP for the last five tasks, respectively. For the ACC, AIA and BWT, BDNet outperforms the second place by 0.048, 0.046 and 0.022, respectively. In a word, BDNet exhibits better memory and generalization performance than existing methods for long task sequences.

Table 6 reports the CTR results of Scenario 6, consisting of five datasets with larger domain gap. Under this scenario, BDNet does not perform well, achieving

three top-2 position and a few top-3 position among all the 18 mAP metrics. The main reason is that, for the CTR scenario with large domain gap, BDNet does not design subtle knowledge distillation scheme, but merely uses the inter-task Banzhaf distillation, which attempts to maintain the invariance of the interaction intensity of each token. For Scenario 6, this attempt may jeopardize sparsely distributed token clues of images or texts. In future work, we shall consider using the affinity correction strategy of DKR [6], and design fine-grained knowledge distillation to improve BDNet.

4.4 Ablation Study on Main Components

Table 7a ablates important components of BDNet: the IETE’s prompts, CLD module and CBD module. According to the result, the performance gradually improves by adding the CBD module and prompt learning scheme. Compared to the base model (w/ CLD, w/o CBD, row 2), BDNet improves by 0.19, 0.13 and 0.11 on the average mAP for Tasks 1–3, and 0.15, 0.16, 0.07 on the ACC, AIA and BWT, respectively. This shows that each module contributes to the improvement of model performance. Table 7b compares different hyperparameter values for $\alpha_1, \dots, \beta$, which supports the hyperparameter selection of Table 3. In particular, the results in the top block of Table 7b indicate that the CLD module serves as the backbone of BDNet, while the CBD module provides satisfactory increment towards the final model.

4.5 Ablation Study on the CBD Module

Let us investigate how the CBD module works towards the CTR task and alleviates the forgetting.

4.5.1 Token similarity visualization

We visualize the textual attention and token similarity of BDNet and BDNet w/o CBD on Scenario 1 in Fig. 3. According to Fig. 3a, BDNet produces three large attention values on **plane**, **runway**, **buildings** (greater than 0.7), which characterizes the main semantic of the

Table 7: Ablation study and hyperparameter selection of BDNet. Gray row denotes our final selection.

(a) Ablation study of BDNet on Scenario 1. Top block denotes results ablating CLD, CBD module or prompt, middle block denotes results remaining one loss among the CBD and CLD modules, and bottom block denotes results ablating one loss among the CBD and CLD modules.

IETE		CLD		CBD		Task1			Task2			Task3			ACC↑	AIA↑	BWT↑
Prompt	Non-prompt	$\mathcal{L}_{CL}$	$\mathcal{L}_{CD}$	$\mathcal{L}_{Banzhaf}$	$\mathcal{L}_{BD}$	I2T	T2I	AVER	I2T	T2I	AVER	I2T	T2I	AVER			
$\times$	$\checkmark$	$\times$	$\times$	$\checkmark$	$\checkmark$	0.0917	0.3363	0.2140	0.0853	0.2886	0.1820	0.0803	0.2003	0.1403	0.1445	0.1762	-0.0671
$\times$	$\checkmark$	$\checkmark$	$\checkmark$	$\times$	$\times$	0.6497	0.6059	0.6278	0.7475	0.7348	0.7412	0.7118	0.7499	0.7305	0.6987	0.6778	-0.0865
$\times$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	<u>0.6689</u>	<u>0.6542</u>	<u>0.6615</u>	0.7835	<u>0.8102</u>	0.7968	<u>0.8025</u>	<u>0.7816</u>	<u>0.7920</u>	<u>0.6998</u>	0.6781	<u>-0.0218</u>
$\checkmark$	$\times$	$\times$	$\times$	$\checkmark$	$\checkmark$	0.3126	0.2633	0.2880	0.3667	0.2436	0.3052	0.2930	0.1933	0.2432	0.2788	0.2567	-0.0517
$\checkmark$	$\times$	$\checkmark$	$\checkmark$	$\times$	$\times$	0.6653	0.6336	0.6494	<u>0.7941</u>	0.8085	<u>0.8013</u>	0.7734	0.7746	0.7740	0.7416	<u>0.7907</u>	-0.0333
$\checkmark$	$\times$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	0.8511	0.7922	0.8216	0.9029	0.8467	0.8748	0.8109	0.8712	0.8410	0.8458	0.8329	-0.0116
$\checkmark$	$\times$	$\checkmark$	$\times$	$\times$	$\times$	0.5735	0.5556	0.5648	0.6283	0.6111	0.6197	0.6667	0.6832	0.6750	0.6190	0.6000	-0.0438
$\checkmark$	$\times$	$\times$	$\checkmark$	$\times$	$\times$	0.1460	0.1143	0.1302	0.1079	0.1746	0.1413	0.9050	0.2000	0.1953	0.1759	0.2514	-0.0578
$\checkmark$	$\times$	$\times$	$\times$	$\checkmark$	$\times$	0.1367	0.1001	0.1184	0.0804	0.1537	0.1171	0.1789	0.1726	0.1758	0.1371	0.1203	-0.6270
$\checkmark$	$\times$	$\times$	$\times$	$\times$	$\checkmark$	0.1326	0.0935	0.1131	0.0981	0.1627	0.1304	0.1832	0.1701	0.1767	0.1401	0.1238	-0.0635
$\checkmark$	$\times$	$\times$	$\checkmark$	$\checkmark$	$\checkmark$	0.3688	0.2840	0.3263	0.3765	0.2647	0.3206	0.3325	0.2119	0.2722	0.3064	0.2846	-0.0543
$\checkmark$	$\times$	$\checkmark$	$\times$	$\checkmark$	$\checkmark$	0.8173	0.7728	0.7950	0.8921	0.8574	0.8747	0.8047	0.8070	0.8058	0.8252	0.8140	-0.0233
$\checkmark$	$\times$	$\checkmark$	$\checkmark$	$\times$	$\checkmark$	0.8307	0.7412	0.7860	0.9070	0.8645	0.8858	0.7858	0.8296	0.8073	0.8264	0.8129	-0.0208
$\checkmark$	$\times$	$\checkmark$	$\checkmark$	$\checkmark$	$\times$	0.6571	0.6857	0.6715	0.7619	0.7810	0.7715	0.7444	0.7333	0.7389	0.7886	0.7593	-0.0324

(b) Hyperparameter selection of BDNet on Scenario 1.

α_1	α_2	α	β	Task1			Task2			Task3			ACC↑	AIA↑	BWT↑
				I2T	T2I	AVER	I2T	T2I	AVER	I2T	T2I	AVER			
0.5				0.8057	0.7606	0.7832	0.9020	0.8781	0.8901	0.8074	0.8501	0.8287	0.8365	0.8229	-0.0144
1	1	0.1	0.3	0.8511	0.7922	0.8216	0.9029	0.8467	0.8748	0.8109	0.8712	0.8410	0.8458	0.8329	-0.0116
1.5				0.8186	0.7664	0.7925	0.8464	0.8559	0.8512	0.9183	0.8602	0.8397	0.8278	0.8258	-0.0197
	0.5			0.7985	0.7631	0.7805	0.8982	0.8804	0.8893	0.8093	0.8334	0.8214	0.8304	0.8117	-0.156
1	1	0.1	0.3	0.8511	0.7922	0.8216	0.9029	0.8467	0.8748	0.8109	0.8712	0.8410	0.8458	0.8329	-0.0116
	1.5			0.7975	0.7565	0.7770	0.9076	0.8497	0.8737	0.8155	0.8334	0.8244	0.8250	0.8199	-0.0222
		0.05		0.7839	0.7400	0.7619	0.8875	0.8680	0.8778	0.8056	0.8191	0.8123	0.8173	0.8085	-0.0269
1	1	0.1	0.3	0.8511	0.7922	0.8216	0.9029	0.8467	0.8748	0.8109	0.8712	0.8410	0.8458	0.8329	-0.0116
		0.2		0.8003	0.7581	0.7792	0.8922	0.8712	0.8817	0.7876	0.8378	0.8127	0.8245	0.8134	-0.0234
			0.1	0.7994	0.7448	0.7721	0.8082	0.7528	0.7805	0.7918	0.8292	0.8105	0.7877	0.7809	-0.0208
1	1	0.1	0.3	0.8511	0.7922	0.8216	0.9029	0.8467	0.8748	0.8109	0.8712	0.8410	0.8458	0.8329	-0.0116
			0.5	0.7910	0.7487	0.7698	0.7843	0.7561	0.7702	0.8143	0.8028	0.8085	0.7828	0.8021	-0.0264

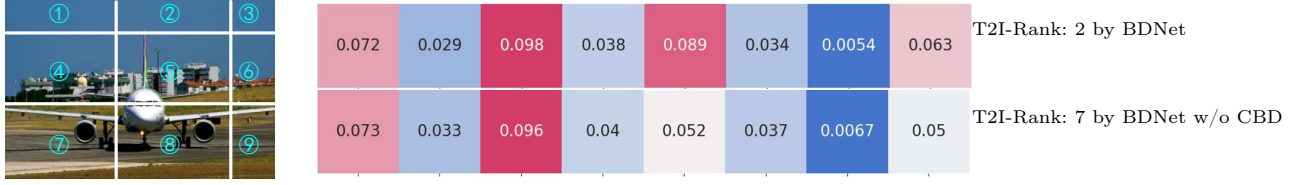
image and achieves rank **2** in text-to-image retrieval, while BDNet w/o CBD produces two large attention values on **plane**, **runway**, leaving small attention over key component: **buildings** of the image and resulting rank **7** in text-to-image retrieval. Let us further examine the token similarity matrix of two models in Figs. 3b and 3c. Thanks to the token-level feature aggregation, BDNet allocates large similarity for **plane**, **runway**, **background** over correct regions, while BDNet w/o CBD fails.

In Fig. 4a, BDNet produces three large attention values on **boat**, **floating**, **water**, which characterizes the main semantic of the image, while BDNet w/o CBD produces one large attention values on **water**, leaving small attention over key component: **boat** of the image. By analyzing the similarity matrix in Fig. 4b and 4c, BDNet allocates large similarity to boat, floating, water and their

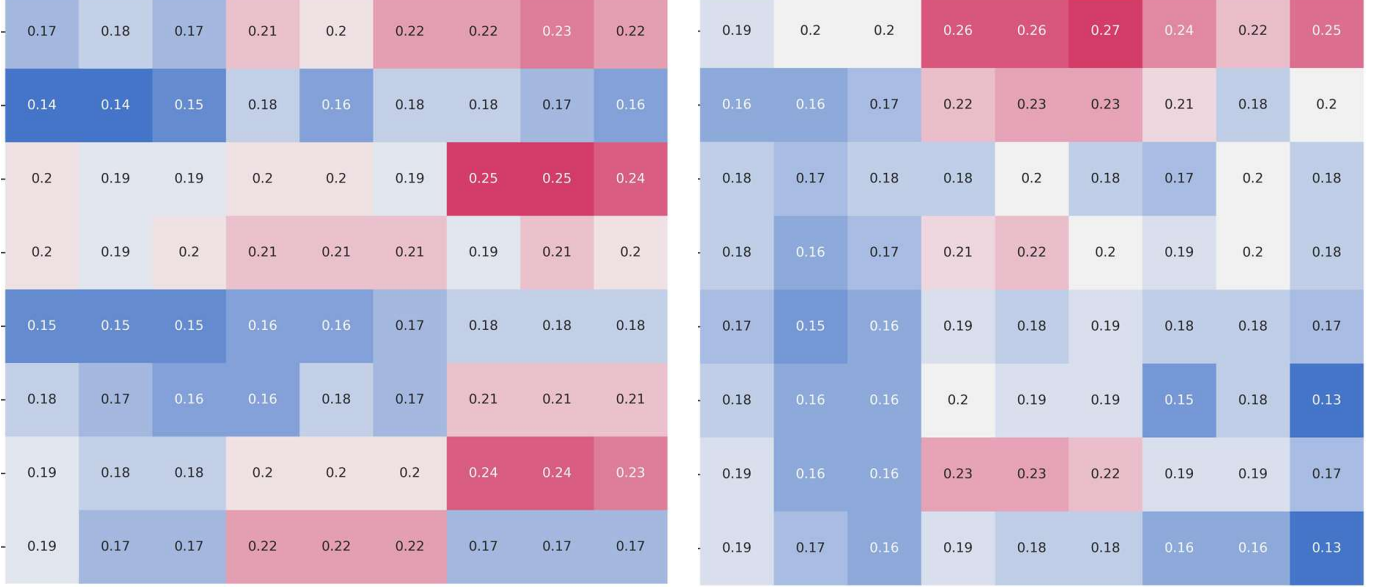
corresponding semantic regions in the image. In contrast, the similarity assignment of BDNet w/o CBD is less accurate. Another visualization result is given in Fig. 5. According to Fig. 5a, BDNet produces three large attention values on **trains**, **inside**, **station**, while BDNet w/o CBD produces one large attention values on **station**, leaving smaller attention weights to the key word **trains**. Further analyzing the similarity matrices in Figs. 4 and 5, we can find that BDNet obtains a more accurate similarity assignment, allocating larger similarity values to **trains**, **station** and **inside**, while BDNet w/o CBD fails.

4.5.2 Text-to-image retrieval ranking

Figs. 6–9 conduct experiments on the antiforgetting ability on CTR ranking results, by using the final model Θ^3



(a) Test image (left) and text attention of BDNet (right top) and BDNet w/o CBD (right bottom), given by the last layer of BERT



(b) Token similarity matrix by BDNet

(c) Token similarity matrix by BDNet w/o CBD

Fig. 3: Visualization of text attention (a) and token similarity (b), (c) on an image-text pair: **plane on runway with buildings in the background**. Each column of (b) and (c) corresponds to average similarity over regions of test image (a).

and the first model Θ^1 over $\mathcal{D}^1$ of Scenario 1 with two models: BDNet and BDNet w/o CBD.

Fig. 6 retrievals images from the query text: **an aeromexico jet taxiing along a runway**. As we see, BDNet w/o CBD (Θ^3) suffers from the forgetting by ranking the ground-truth image at the third position, while BDNet handles the forgetting well, thanks to the powerful token group modeling towards the relation between the jet and runway.

Fig. 7 retrievals images from the query text: **an airliner has taken off behind one which is on the ground**. The ground-truth retrieval image consists of a small airliner and a large airliner, which increases the retrieval difficulty. Surprisingly, BDNet retrieves it correctly at the second position by the final model Θ^3 though it ranks at the seventh position by the first model Θ^1 . In contrast, after removing the CBD module, the model fails to rank the image in front.

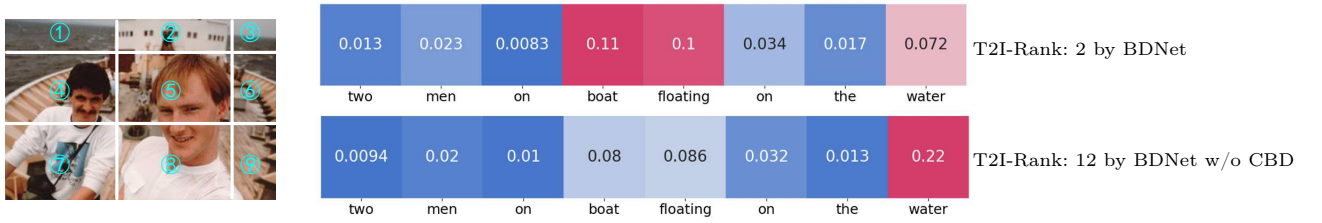
Fig. 8 retrievals images from the query text: **a white car driving past two domed buildings**, which is a similar case to Fig. 7, with a smaller white car on the road at the bottom of the ground-truth image. In this case, the

final model Θ^3 of BDNet ranks it at the third position, better than BDNet w/o CBD.

Fig. 9 retrievals images from the query text: **a man waits on the platform as the light rail train approaches**. Readers can see that this retrieval job is difficult since the man in the ground-truth image appears partially. Yet, both the first model Θ^1 and the final model Θ^3 of BDNet rank it with satisfactory order (second and third position). Those examples indicate that, when images contain multiple objects with different sizes, BDNet captures multi-granular semantics based on its group modeling ability induced from the Banzhaf value.

4.5.3 Image-to-text retrieval ranking

Fig. 10 gives image-to-text retrieval ranking results for three examples of Scenario 1, with the ground-truth texts marked by red boxes. Those three query images, indexed by 49, 93, 99 in Scenario 1, involve high confusing text descriptions. Besides, each of image involves the relative position or relation of two objects. For instance, the second image is the example of Fig. 9, which cannot be retrieved in top-10 by BDNet w/o CBD. As we see, although those three query images are not ranked in the



(a) Test image (left) and text attention of BDNet (right top) and BDNet w/o CBD (right bottom), given by the last layer of BERT

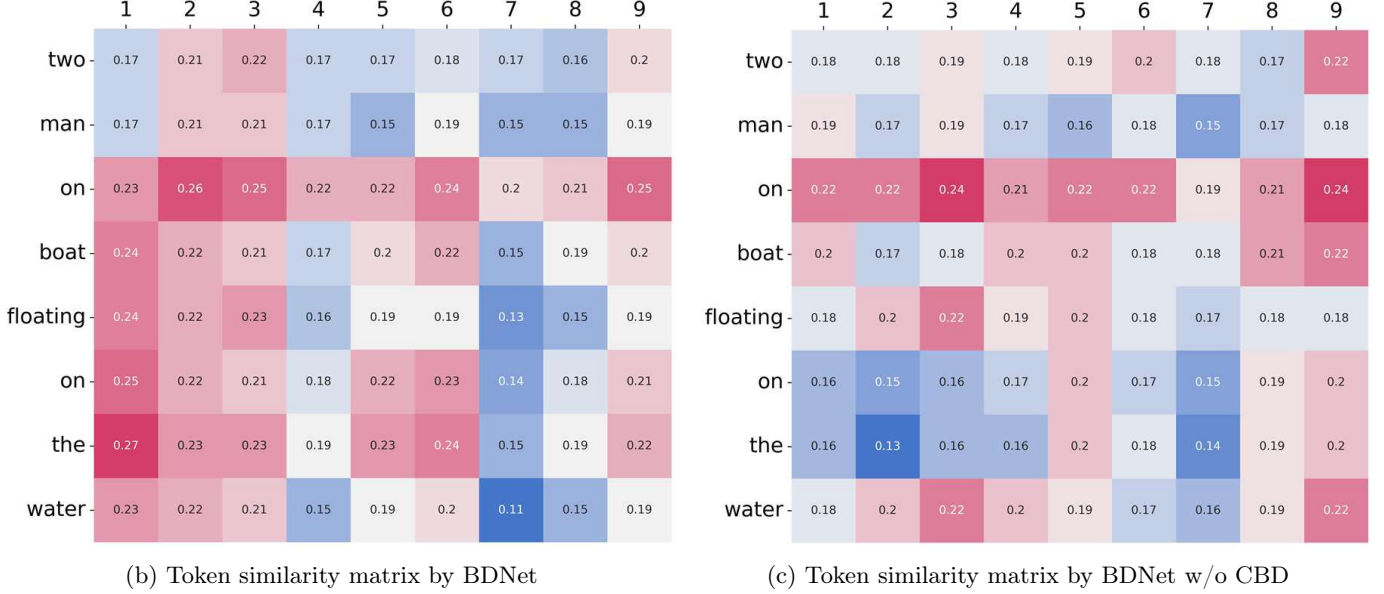


Fig. 4: Visualization of text attention (a) and token similarity (b), (c) on an image-text pair: **two men on boat floating on the water**. Each column of (b) and (c) corresponds to average similarity over regions of test image (a).

top-3 order by any of the methods, BDNet models the complicated relation of objects in the three images well by using the CBD module.

4.6 Computation Overhead Comparison

Table 8 lists the number of parameters and GFLOPs of each module in BDNet, including image encoder, text encoder, prompt module and Banzhaf module. According to the table, the complexity of BDNet is 5.13 GFLOPs and the number of parameters is 84.24M, while the Banzhaf module consists of linear and self-attentive layers and has small complexity. Also, the prompt module brings only a minimal growth in parameters, which makes BDNet lightweight enough. In addition, the Banzhaf module has good portability and can be embedded into any image-text retrieval network to improve the performance without increasing the complexity of the model.

5 Conclusions

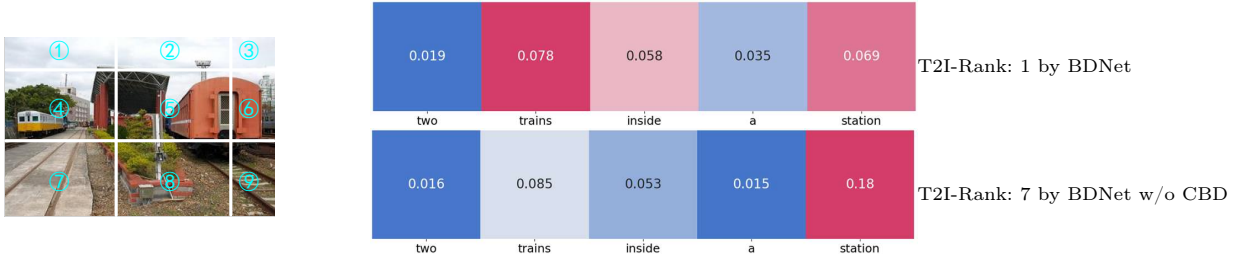
Towards the cross-modal semantic discrepancy and the cross-task catastrophic forgetting of knowledge incurred from continual image-text retrieval (CITR), we propose

Table 8: Parameter number and GFLOPs of different modules of BDNet.

Modules	ParaNum(M)	GFLOPs
Image Encoder	59.02	3.11
Text Encoder	25.07	1.93
Image Prompt	0.016	0
Text Prompt	0.012	0
Banzhaf module	0.13	0.09
Total	84.24	5.13

BanZhaf Distillation Network (BDNet). Based on the tokenized configuration of transformers and motivated from the analog of data valuation to token contribution, BDNet generalizes single Banzhaf value to Banzhaf value, and formulates the token-level similarity of an image and a text as a coalitional game to preserve fine-grained cross-modal knowledge. Furthermore, BDNet distills the token similarity using replay data to memorize cross-task knowledge. Experiments on both single-set and multi-set scenarios demonstrate the effectiveness over traditional methods.

The shortcoming of BDNet is twofold. First, high efficient evaluation of BanZhaf values is necessary. BDNet



(a) Test image (left) and text attention of BDNet (right top) and BDNet w/o CBD (right bottom), given by the last layer of BERT

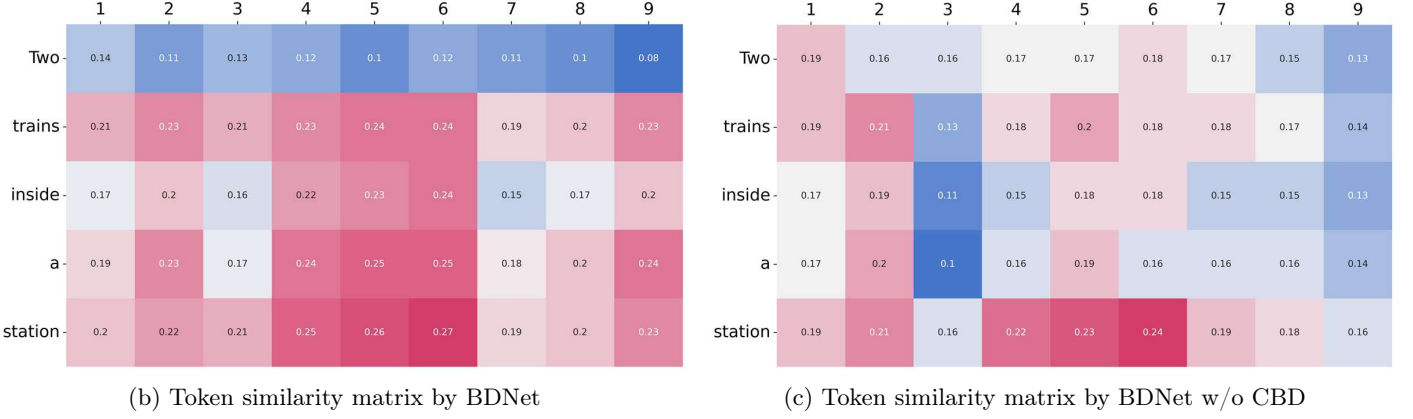


Fig. 5: Visualization of text attention (b) and token similarity (c), (d) on an image-text pair: **Two trains inside a station**. Each column of (c) and (d) corresponds to average similarity over regions of (a).



Fig. 6: Text-to-image retrieval ranking results by the final model Θ^3 and the first model Θ^1 on $\mathcal{D}^1$ for Scenario 1. Query text: **an aeromexico jet taxiing along a runway**. Red boxes denote ground-truth images. Left to right denotes the top-8 ranking images. While BDNet w/o CBD (rows 1–2) indicates serious forgetting on the text-to-image correspondence, BDNet (rows 3–4) alleviates the forgetting by characterizing the scene text with token group modeling.



Fig. 7: Text-to-image retrieval ranking results by the final model Θ^3 and the first model Θ^1 on $\mathcal{D}^1$ for Scenario 1. Query text: **an airliner has taken off behind one which is on the ground**. Red boxes denote ground-truth images. Left to right denotes the top-8 ranking images. This example involves two airplanes and their relative position, and BDNet (rows 3–4) characterizes two kinds of objects due to its multi-granularity modeling ability by the CBD module.

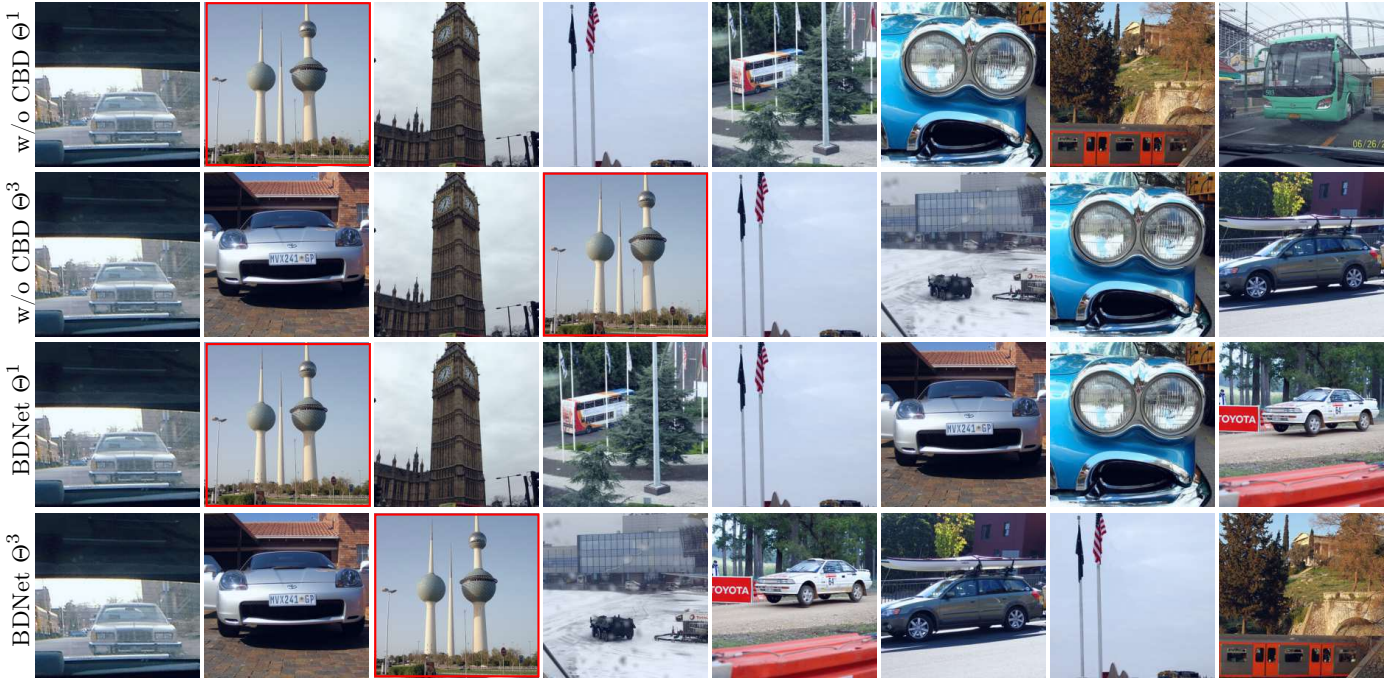


Fig. 8: Text-to-image retrieval ranking results by the final model Θ^3 and the first model Θ^1 on $\mathcal{D}^1$ for Scenario 1. Query text: **a white car driving past two domed buildings**. Red boxes denote ground-truth images. Left to right denotes the top-8 ranking images. This example involves large scale buildings and tiny scale white car, and BDNet (rows 3–4) characterizes two kinds of objects due to its multi-granularity modeling ability induced from the CBD module.

Acknowledgment

The research work is supported by the National Natural Science Foundation of China under Grant 62441232, and the Beijing Natural Science Foundation under Grant L257022.

References

- [1] Li, W., Gao, B.-B., Xia, B., Wang, J., Liu, J., Liu, Y., Wang, C., Zheng, F.: Cross-Modal Alternating Learning With Task-Aware Representations for Continual Learning. *IEEE Transactions on Multimedia* **26**, 5911–5924 (2024)
- [2] Lee, K.-H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: *Proceedings of the European Conference on Computer Vision*, pp. 212–228 (2018)
- [3] Chen, H., Ding, G., Liu, X., Lin, Z., Liu, J., Han, J.: IMRAM: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12652–12660 (2020)
- [4] Diao, H., Zhang, Y., Ma, L., Lu, H.: Similarity reasoning and filtration for image-text matching. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 1218–1226 (2021)
- [5] Ni, Z., Wei, L., Tang, S., Zhuang, Y., Tian, Q.: Continual Vision-Language Representation Learning with Off-Diagonal Information. In: *International Conference on Machine Learning*, vol. 202, pp. 26129–26149. PMLR, ??? (2023)
- [6] Cui, Z., Peng, Y., Wang, X., Zhu, M., Zhou, J.: Continual vision-language retrieval via dynamic knowledge rectification. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 11704–11712 (2024)
- [7] Dong, J., Li, X., Xu, C., Yang, X., Yang, G., Wang, X., Wang, M.: Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(8), 4065–4080 (2022)
- [8] Peng, Y., Qi, J., Ye, Z., Zhuo, Y.: Hierarchical visual-textual knowledge distillation for life-long correlation learning. *International Journal of Computer Vision* **129**(4), 921–941 (2021)
- [9] Song, G., Tan, X.: Real-world cross-modal retrieval via sequential learning. *IEEE Transactions on Multimedia* **23**, 1708–1721 (2021)
- [10] Yan, W., Wang, Y., Lin, W., Guo, Z., Zhao, Z., Jin, T.: Low-rank Prompt Interaction for Continual Vision-Language Retrieval. In: *Proceedings of the ACM International Conference on Multimedia*, pp. 8257–8266 (2024)
- [11] Wang, J.T., Jia, R.: Data Banzhaf: A robust data valuation framework for machine learning. In: *Proceedings of the International Conference on Artificial Intelligence and Statistics*, vol. 206, pp. 6388–6421 (2023)
- [12] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S.: An image is worth 16x16 words: Transformers for image recognition at scale. In: *International Conference on Learning Representation*, pp. 1–21 (2020)
- [13] Zhu, H., Zhang, C., Wei, Y., Huang, S., Zhao, Y.: ESA: External space attention aggregation for image-text retrieval. *IEEE Transactions on Circuits and Systems for Video Technology* **33**(10), 6131–6143 (2023)
- [14] Liu, C., Zhang, Y., Wang, H., Chen, W., Wang, F., Huang, Y., Shen, Y.-D., Wang, L.: Efficient token-guided image-text retrieval with consistent multi-modal contrastive training. *IEEE Transactions on Image Processing* **32**, 3622–3633 (2023)
- [15] Jin, P., Li, H., Yuan, L., Yan, S., Chen, J.: Hierarchical Banzhaf Interaction for General Video-Language Representation Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **47**(3), 2125–2139 (2025)
- [16] Ferguson, T.S.: *A Course in Game Theory*. Singapore, New York, NY, USA (2020)
- [17] Jia, R., Wu, F., Sun, X., Xu, J., Dao, D., Kailkhura, B., Zhang, C., Li, B., Song, D.: Scalability vs. utility: Do we have to sacrifice one for the other in data importance quantification? In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8235–8243 (2021)
- [18] Jia, R., Dao, D., Wang, B., Hubis, F.A., Hynes, N., Gürel, N.M., Li, B., Zhang, C., Song, D., Spanos, C.J.: Towards efficient data valuation based on the Shapley value. In: *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 1167–1176 (2019)
- [19] Peng, Y., Huang, X., Zhao, Y.: An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges. *IEEE Transactions on Circuits and Systems for Video Technology* **28**(9), 2372–2385 (2018)

- [20] Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G.R.G., Levy, R., Vasconcelos, N.: A new approach to cross-modal multimedia retrieval. In: *Proceedings of the ACM International Conference on Multimedia*, pp. 251–260 (2010)
- [21] Costa Pereira, J., Coviello, E., Doyle, G., Rasiwasia, N., Lanckriet, G.R.G., Levy, R., Vasconcelos, N.: On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **36**(3), 521–535 (2014)
- [22] Rasiwasia, N., Mahajan, D., Mahadevan, V., Aggarwal, G.: Cluster canonical correlation analysis. In: *Proceedings of the International Conference on Artificial Intelligence and Statistics*, vol. 33, pp. 823–831 (2014)
- [23] Ranjan, V., Rasiwasia, N., Jawahar, C.V.: Multi-label cross-modal retrieval. In: *IEEE International Conference on Computer Vision*, pp. 4094–4102 (2015)
- [24] Gu, J., Cai, J., Joty, S., Niu, L., Wang, G.: Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7181–7189 (2018)
- [25] Zhang, Y., Lu, H.: Deep cross-modal projection learning for image-text matching. In: *Proceedings of the European Conference on Computer Vision*, pp. 707–723 (2018)
- [26] Zhen, L., Hu, P., Wang, X., Peng, D.: Deep supervised cross-modal retrieval. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10386–10395 (2019)
- [27] Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: DeViSE: a deep visual-semantic embedding model. In: *Advances in Neural Information Processing Systems*, pp. 2121–2129 (2013)
- [28] Liu, Y., Guo, Y., Bakker, E.M., Lew, M.S.: Learning a recurrent residual fusion network for multimodal matching. In: *IEEE International Conference on Computer Vision*, pp. 4127–4136 (2017)
- [29] Wang, L., Li, Y., Huang, J., Lazebnik, S.: Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41**(2), 394–407 (2019)
- [30] Li, S., Xiao, T., Li, H., Yang, W., Wang, X.: Identity-aware textual-visual matching with latent co-attention. In: *IEEE International Conference on Computer Vision*, pp. 1908–1917 (2017)
- [31] Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5005–5013 (2016)
- [32] Zheng, Z., Zheng, L., Garrett, M., Yang, Y., Xu, M., Shen, Y.-D.: Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications* **16**(2) (2020)
- [33] Jian, Y., Xiao, J., Cao, Y., Khan, A., Zhu, J.: Deep pairwise ranking with multi-label information for cross-modal retrieval. In: *IEEE International Conference on Multimedia and Expo*, pp. 1810–1815 (2019)
- [34] Zhao, S., Jia, Z., Chen, H., Li, L., Ding, G., Keutzer, K.: PDANet: Polarity-consistent deep attention network for fine-grained visual emotion regression. In: *Proceedings of the ACM International Conference on Multimedia*, pp. 192–201 (2019)
- [35] Cui, Q., Jiang, Q.-Y., Wei, X.-S., Li, W.-J., Yoshie, O.: ExchNet: A unified hashing network for large-scale fine-grained image retrieval. In: *Proceedings of the European Conference on Computer Vision*, pp. 189–205 (2020)
- [36] Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39**(4), 664–676 (2017)
- [37] Li, W., Wang, Y., Su, Y., Li, X., Liu, A.-A., Zhang, Y.: Multi-scale fine-grained alignments for image and sentence matching. *IEEE Transactions on Multimedia* **25**, 543–556 (2023)
- [38] Wang, Y., Su, Y., Li, W., Xiao, J., Li, X., Liu, A.-A.: Dual-path rare content enhancement network for image and text matching. *IEEE Transactions on Circuits and Systems for Video Technology* **33**(10), 6144–6158 (2023)
- [39] Fu, Z., Mao, Z., Song, Y., Zhang, Y.: Learning semantic relationship among instances for image-text matching. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15159–15168 (2023)
- [40] Diao, H., Zhang, Y., Liu, W., Ruan, X., Lu, H.: Plug-and-play regulators for image-text matching. *IEEE Transactions on Image Processing* **32**, 2322–2334 (2023)
- [41] Xiong, G., Meng, M., Zhang, T., Zhang, D., Zhang,

- Y.: Reference-aware adaptive network for image-text matching. *IEEE Transactions on Circuits and Systems for Video Technology* **34**(10), 9678–9691 (2024)
- [42] Ji, Z., Wang, H., Han, J., Pang, Y.: SMAN: Stacked multimodal attention network for cross-modal image-text retrieval. *IEEE Transactions on Cybernetics* **52**(2), 1086–1097 (2022)
- [43] Zhang, K., Mao, Z., Liu, A.-A., Zhang, Y.: Unified adaptive relevance distinguishable attention network for image-text matching. *IEEE Transactions on Multimedia* **25**, 1320–1332 (2023)
- [44] Jin, P., Huang, J., Xiong, P., Tian, S., Liu, C., Ji, X., Yuan, L., Chen, J.: Video-text as game players: Hierarchical banzhaf interaction for cross-modal representation learning. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2472–2482 (2023)
- [45] Wang, K., Herranz, L., Weijer, J.: Continual learning in cross-modal retrieval. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop*, pp. 3623–3633 (2021)
- [46] Yang, R., Wang, S., Zhang, H., Xu, S., Guo, Y., Ye, X., Hou, B., Jiao, L.: Knowledge decomposition and replay: A novel cross-modal image-text retrieval continual learning method. In: *Proceedings of the ACM International Conference on Multimedia*, pp. 6510–6519 (2023)
- [47] Zhu, X., Jiang, Z., Wu, K., Shi, J., Zheng, Y.: Life-long Histopathology Whole Slide Image Retrieval via Distance Consistency Rehearsal. In: *Medical Image Computing and Computer Assisted Intervention*, pp. 274–284 (2024)
- [48] Yang, R., Wang, S., Gu, Y., Wang, J., Sun, Y., Zhang, H., Liao, Y., Jiao, L.: Continual learning for cross-modal image-text retrieval based on domain-selective attention. *Pattern Recognition* **149**, 110273–113 (2024)
- [49] Zhang, H., Yang, Y., Qi, F., Qian, S., Xu, C.: C2MR: Continual cross-modal retrieval for streaming multi-modal data. In: *Proceedings of the ACM International Conference on Multimedia*, pp. 8963–8974 (2023)
- [50] Liu, W., Zhu, F., Wei, L., Tian, Q.: C-CLIP: Multimodal continual learning for vision-language model. In: *International Conference on Learning Representation*, pp. 1–17 (2025)
- [51] Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 1715–1725 (2016)
- [52] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 4171–4186 (2019)
- [53] Wang, Z., Zhang, Z., Lee, C.-Y., Zhang, H., Sun, R., Ren, X., Su, G., Perot, V., Dy, J., Pfister, T.: Learning to prompt for continual learning. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 139–149 (2022)
- [54] Roubens, M., Grabisch, M.: An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory* **28**(4), 547–565 (1999)
- [55] Li, Z., Hoiem, D.: Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **40**(12), 2935–2947 (2018)
- [56] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., *et al.*: Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences* **114**(13), 3521–3526 (2017)
- [57] Del Chiaro, R., Twardowski, B., Bagdanov, A., Weijer, J.: RATT: Recurrent attention to transient tasks for continual image captioning. In: *Advances in Neural Information Processing Systems*, vol. 33, pp. 16736–16748 (2020)
- [58] Rashtchian, C., Young, P., Hodosh, M., Hockenmaier, J.: Collecting image annotations using Amazon’s Mechanical Turk. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics Workshop*, pp. 139–147 (2010)
- [59] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: *Proceedings of the European Conference on Computer Vision*, pp. 740–755 (2014)
- [60] Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. In: *IEEE International Conference on Computer Vision*, pp. 2641–2649 (2015)
- [61] Grubinger, M., Clough, P., Müller, H., Deselaers, T.: The IAPR TC12 Benchmark: A New Evaluation

Resource for Visual Information Systems. Workshop Ontoimage, 1–11 (2006)

- [62] Lin, J., Men, R., Yang, A., Zhou, C., Zhang, Y., Wang, P., Zhou, J., Tang, J., Yang, H.: M6: Multi-Modality-to-Multi-Modality Multitask Megatransformer for Unified Pretraining. In: ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 3251–3261 (2021)
- [63] Lu, X., Wang, B., Zheng, X., Li, X.: Exploring Models and Data for Remote Sensing Image Caption Generation. *IEEE Transactions on Geoscience and Remote Sensing* **56**(4), 2183–2195 (2018)
- [64] Dong, S., Hong, X., Tao, X., Chang, X., Wei, X., Gong, Y.: Few-Shot Class-Incremental Learning via Relation Knowledge Distillation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 1255–1263 (2021)
- [65] Kang, M., Park, J., Han, B.: Class-Incremental Learning by Knowledge Distillation with Adaptive Feature Consolidation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16050–16059 (2022)