SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

May 2026, Vol. 69, Iss. 5, 152104:1–152104:17
https://doi.org/10.1007/s11432-024-4845-2

Improving image-text alignment with an optimal feature sub-space-aware similarity learning framework

Kun ZHANG[1,2†], Jingyu LI[3†], Zhe LI[4], Huatian ZHANG[5], Lei ZHANG[5], Zhendong MAO[3,4] & Yongdong ZHANG[3,5*]

[1]*Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou 215123, China*
[2]*School of Biomedical Engineering, University of Science and Technology of China, Hefei 230022, China*
[3]*Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230022, China*
[4]*School of Cyberspace Science and Technology, University of Science and Technology of China, Hefei 230022, China*
[5]*School of Information Science and Technology, University of Science and Technology of China, Hefei 230022, China*

Received 6 November 2024/Revised 7 February 2025/Accepted 23 March 2025/Published online 17 April 2026

Abstract Image-text alignment serves as a fundamental cross-modal research topic to bridge vision and language. Its key challenge lies in accurately measuring the similarity of these two heterogeneous modalities. For visual and textual features, most existing methods leverage cosine or Euclidean distance to measure similarity, where the modality features are directly examined in the whole representation space. However, we discover that partial local dimensions, forming sub-spaces with the potential semantic representation tendency, contain more important semantic measurement information. Thus, we argue that existing methods fail to focus on the finer alignment of critical sub-spaces composed of partial dimensions, leading to limited and inaccurate similarity learning. To address this problem, we propose a novel optimal feature sub-space-aware similarity learning framework (OPEN), which takes a forward step to focus on the sub-space composed of local dimensions within modality representations, enabling more subtle semantic alignment and similarity measurement. Specifically, we first construct hierarchical sub-space-aware patterns for learning similarity, i.e., the sub-space comprised of different sizes of local dimensions. Then, for the optimality of the OPEN, there are two new aspects: (1) optimal sub-space-aware patterns, where we reveal which size-level of local dimensions in the sub-space pattern can achieve the optimal performance gains with maximum probability; (2) optimal combined sub-space-aware patterns, in which we mine the optimal complementarities of different size-level patterns. The proposed OPEN enjoys the merit of plug-and-play, and we extensively experiment with it on typical cross-modal alignment paradigms and datasets. OPEN offers consistent and significant performance improvements across different settings, verifying its superiority for simplicity, generality, and effectiveness.

Keywords image-text alignment, cross-modal semantic measurement, sub-space-aware similarity learning

Citation Zhang K, Li J Y, Li Z, et al. Improving image-text alignment with an optimal feature sub-space-aware similarity learning framework. Sci China Inf Sci, 2026, 69(5): 152104, https://doi.org/10.1007/s11432-024-4845-2

1 Introduction

Associating the most prevalent modalities of vision and language is significant for artificial intelligence in understanding our real world. Image-text alignment is a fundamental research topic to bridge these two heterogeneous modalities, which can be used to search aligned texts by a given visual image and vice versa. Although the past few years have witnessed a surge of studies, it remains challenging to accurately measure the similarity between images and texts.

During the last decades, existing image-text alignment methods can be mainly classified into two lines, i.e., one-to-one (O2O) and many-to-many (M2M). The O2O method maps each image or text as a holistic representation, and then learns the whole semantic alignment to measure image-text similarity [1]. Representative studies, such as VSE++ [2], VSE∞ [3], and CLIP [4], are to constrain the similarity of aligned image-text pairs to be higher than that of the most similar misaligned one. The M2M method extracts the fragmental representations of salient image regions (or patches) and textual words, and then learns fine-grained semantic alignment to infer the overall image-text similarity [5]. One of the most typical studies is SCAN [6], which resorts to the stacked cross-attention to discover all possible region-word alignments, inspiring a series of attention-based studies [7–10]. Commonly, for the textual and visual features in whether O2O or M2M, most existing methods [3, 11–14] typically use the cosine

* Corresponding author (email: zhyd73@ustc.edu.cn)
† These authors contributed equally to this work.

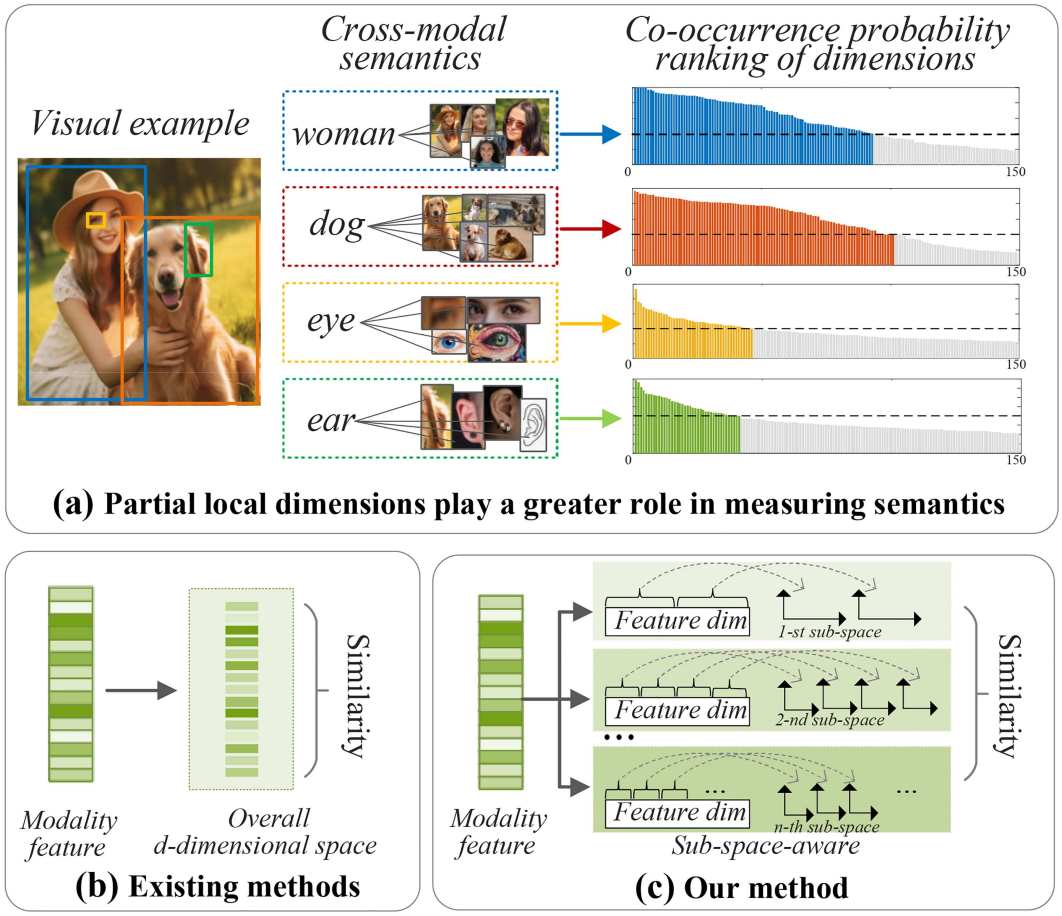

Figure 1 (Color online) (a) The statistical experiments to investigate the role of local dimensions in measuring cross-modal semantic similarity. We first collect all word-region pairs of certain semantics (e.g., woman, dog, eye, and ear) in the Flickr30K training set. Considering that the cross-modal semantic similarity of a word-region pair is reflected by all feature dimensions, we select the top-K ($K = 50$) contributing dimensions for each pair, and count the co-occurrence probabilities of these dimensions in all pairs with the same semantics, under the state-of-the-art model [14]. The greater the co-occurrence probability of a dimension, the more important the dimension is for measuring the semantics. Dimensions with a co-occurrence probability greater than 0.4 are marked with colors, where we can find that only partial dimensions play a greater role in measuring semantics, and the number of partial dimensions also varies with the corresponding semantics. (b) Most existing methods typically ignore this property to directly measure similarity in the overall d-dimensional space. (c) Our method captures the varied importance of the sub-space composed of partial dimensions of different scales, towards finer and more accurate similarity measurement.

or Euclidean distance to measure their similarity, as shown in Figure 1(b), where the modality feature is directly measured in the overall d-dimensional representation space.

However, as an interesting finding in Figure 1(a), a portion of local dimensions in the cross-modal features plays a greater role in measuring semantics. That is, by conducting statistical experiments, we count the co-occurrence probabilities of the dimensions that have a prominent contribution to the similarity measurement. We discover that partial dimensions have larger co-occurrence probabilities than other dimensions (e.g., the colored dimensions represent the parts with a co-occurrence probability greater than 0.4). In other words, in the whole space spanned by d-dimensions, partial dimensions, constituting a sub-space, are more inclined to represent potential semantics, which contain more important semantic measurement information. Moreover, as shown in Figure 1(a), as the visual semantic complexity increases (e.g., the 'woman' region requires richer detailed characteristics than the region 'ear' to describe), the number of partial dimensions with greater representational tendency also increases. For example, the number of dimensions with colors in cross-modal semantics 'woman' and 'dog' is significantly greater than that in 'eye' and 'ear'. Theoretically, it is consistent with the fact that the more dimensions there are, the more powerful the representation ability is.

In contrast, existing image-text alignment methods typically ignore this inherent property, i.e., the sub-space composed of partial dimensions of different scales plays a more important semantic measurement role. Consequently, existing methods can neither adaptively and dynamically focus on finer alignment of critical sub-spaces, nor explore the optimal measurement of sub-spaces with varied dimension scales, which may result in limited and inaccurate

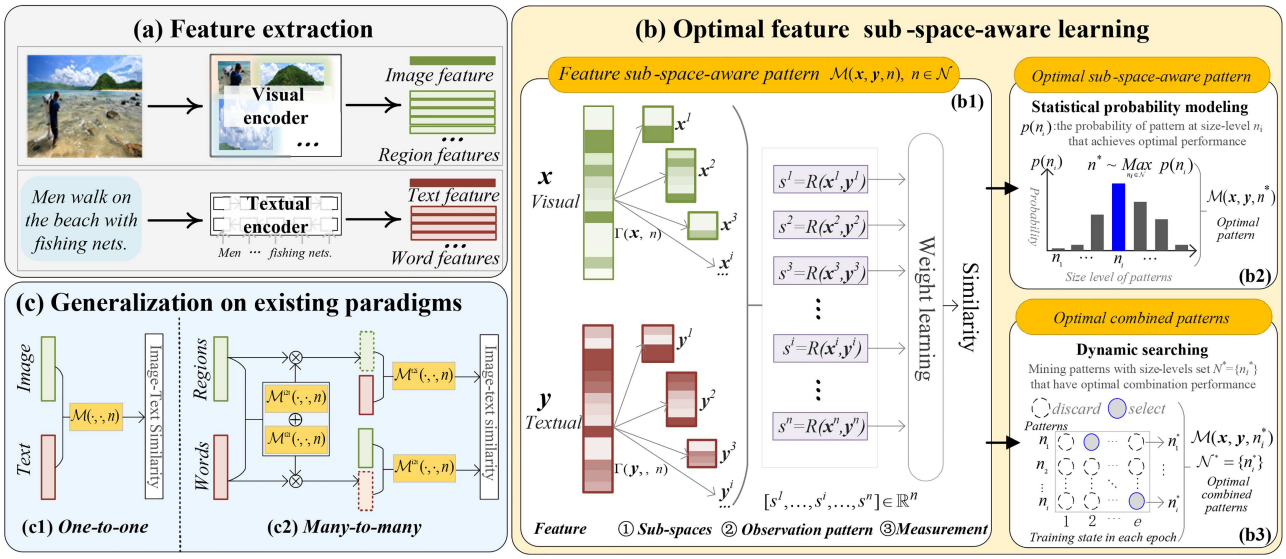

Figure 2 (Color online) An overview of our OPEN framework. (a) Extract visual and textual features. (b) Construct the sub-space-aware pattern $\mathcal{M}(X,Y,n)$, which is comprised of local dimensions at the n size-level, forming the observation pattern for similarity measurement, with optimality: (1) an optimal pattern, mined by modeling the statistical probability of each size level that achieves optimal performance; (2) optimal combined patterns, mined by dynamically searching for patterns at different size levels to model their optimal complementarities. (c) Generalize OPEN to existing paradigms of image-text alignment.

cross-modal semantic similarity learning.

To address the above issue, we propose a novel optimal feature sub-space-aware similarity learning framework (OPEN) to divide the prior whole space into internal feature sub-spaces, where each sub-space contains partial dimensions to form a local observation window for more refined semantic similarity measurement, as depicted in Figure 1(c). Specifically, as shown in Figure 2, the novelty of our OPEN is 2-fold. (1) We propose to reveal which sub-space-aware pattern can bring the optimal performance gains with maximum probability, explored by a parallel learning architecture with patterns at different size-level local sub-spaces. Each pattern is achieved by three steps: designing a local sub-space, obtaining a sub-space-aware pattern, and weighting the similarity measurement. (2) We propose to mine the optimal combined patterns, where such a design allows us to comprehensively measure the similarity of latent semantics in the image-text pairs. Taking the output of each pattern as an optional state in the training process, we derive the optimal state transition equation to dynamically search for their complementarities, yielding a precise similarity measurement.

Our contributions are summarized as follows. (1) We propose a simple yet effective optimal feature sub-space-aware similarity learning framework, which enables the capturing of finer alignment of partial dimensions within the vision-language features. Moreover, we reveal that, for the first time to our knowledge, which size-level local sub-space can achieve the optimal performance gains with maximum probability. (2) We propose a novel optimal combination patterns mining method, which dynamically discovers the optimal complementarities of different size-level sub-space-aware patterns, thereby producing comprehensive and accurate similarity measurements. (3) Our proposed OPEN framework has the merit of plug-and-play, and we propose a way to equip OPEN with the existing two typical image-text matching paradigms. Extensive experiments on various settings (e.g., datasets, paradigms, and local basis designs) verify its effectiveness, where OPEN consistently and significantly brings performance gains, with relative average overall improvements of 3.78% and 1.78% on Flickr30K and MS-COCO, respectively.

2 Related work

In this section, we first introduce two well-established paradigms in image-text alignment: (1) O2O, which typically represents the whole image or the full text as a holistic feature to measure image-text similarity; (2) M2M, which focuses on inferring the overall image-text similarity via all fine-grained word-region similarities. Then, we review the relevant literature on the similarity measurement of cross-modal features.

One-to-one methods. A common solution in this field is to learn a latent embedding space, so that the similarity between the mapped image and text features can be directly measured. Early studies employ multiple neural networks to improve the semantic embeddings of images and texts [11, 15, 16]. Some studies focus on the

aggregation strategy for the holistic presentation of image or text [3,17,18], such as the policy gradient optimization for attention (aggregation) weights. There is another line of studies that focuses on designing delicate optimization functions [2,19], such as the famous hinge-based triplet loss [2]. Recently, there have been some novel optimization designs [20, 21], such as the ladder loss [20] that constrains a continuous relationship of negative samples, and the unified training objective [22] that incorporates and improves the current hubness-aware loss function with momentum contrastive learning.

Many-to-many methods. These studies focus on measuring all similarities between the text words and image regions, aiming to exploit such fine-grained word-region similarities to infer the overall image-text similarity [3,5,11, 23]. Most studies in this field are based on the attention mechanism [6,8,9,24,25]. One of the most typical approaches is the attention-based SCAN [6] that attends to the specific regions/words fragments to discover all latent region-word correspondences. Recent improvements include utilizing local-global and other multi-granularity hierarchical information [26–28], multi-view semantic modeling to solve semantic ambiguity [29], causal reasoning [30], and lightweight inference enhancement [31, 32]. Meanwhile, a line of methods [24, 33, 34] focus on exploiting external information to further enhance the association of images and texts. Representative studies include using scene graphs [34, 35] for relationship modeling to further enhance image-text alignment performance.

Recently, pre-training vision-language models [4, 36–40] based on large-scale image-text data have attracted widespread attention. The representative CLIP [4] has a wide range of application potential and has shown strong performance in image and text tasks in multiple fields, including improving efficiency via LiT-tuning [41], biological microscopy [42], and social image retrieval [43]. Unlike these pre-training studies, our method focuses on the importance of different sub-spaces of visual and text features in the semantic similarity measurement. Moreover, the proposed method can be plug-and-play, and we verify the effectiveness of the similarity measurement method based on the pre-trained backbones.

Similarity learning of cross-modal features. For similarity learning of cross-modal features, most existing methods typically employ the cosine distance [6–8,34], Euclidean distance, or Hamming distance to measure image-text similarity, where the whole feature is directly measured in the semantic space. Compared with these metrics that do not include any learnable parameters, the learnable similarity measurement has received much attention, since it can adaptively and flexibly capture complex patterns from input cross-modal features. In recent years, Liu et al. [24] designed the operation to propagate features on the graph through fully connected layers to obtain similarity, and this operation is used in the later studies [9,44]. Yet this operation essentially adds learnable weights to each dimension. Different from the above work, our work focuses on the varying roles of sub-spaces composed of some local dimensions in the overall semantic representation in similarity measurement. Its goal is to achieve a more refined learning of feature internal sub-spaces and a more comprehensive similarity learning through sub-space observation patterns at different size levels.

3 Method

The overview of the proposed framework is illustrated in Figure 2, which divides the whole feature space into internal sub-spaces, serving as a more subtle local observation basis, to reveal the optimal learning pattern for measuring cross-modal semantic similarity. In this section, we first describe the details of the OPEN framework. Then, the generalization of our proposed OPEN on existing image-text alignment paradigms is introduced in Subsection 3.2. Finally, we present the objective function in Subsection 3.3. Details about feature extraction can be seen in Subsection 3.4.

3.1 Optimal feature sub-space-aware learning

Without loss of generality, we define the input visual instance as $\boldsymbol{x}$, and the textual instance as $\boldsymbol{y}$, where $\boldsymbol{x}$ can be the holistic image in O2O or the fragmental region in M2M (symmetrical for $\boldsymbol{y}$). Our goal is to measure the similarity between $\boldsymbol{x}$ and $\boldsymbol{y}$, under the proposed feature sub-space-aware learning framework, which adaptively learns similarity based on the observation sub-spaces, i.e., composed of local feature dimensions.

As shown in Figure 2, the optimality of our proposed framework includes two novel aspects. (1) Optimal sub-space-aware pattern, aiming to solve the problem: *"Which kind of the sub-space-aware learning pattern is optimal?"*. In fact, the whole feature space can be divided into different numbers of sub-spaces, resulting in varied observation patterns. This is because in different patterns, the sub-space contains varied numbers of partial local dimensions, resulting in different semantic expressive capabilities. Thus, we aim to reveal which learning pattern can achieve optimal performance with maximum probability. (2) Optimal combined sub-space-aware patterns, aiming to solve the problem: *"What sub-space-aware patterns are combinatorially optimal?"*. Considering that

there may be complementary relationships between different sub-space-aware patterns that characterize different abilities, our target is to search for the optimal combination of the local sub-spaces at different size levels, revealing their complementarities for optimal performance. Next, we first introduce the construction of the sub-space-aware pattern, and then elaborate on the optimality of the two aspects in detail.

3.1.1 *Constructing sub-space-aware pattern*

To address the problem that existing methods typically ignore the fact that the sub-space composed of partial dimensions of different scales has more important semantic measurement effects, our learning framework proposes to focus on the internal feature sub-space, through which the cross-modal alignment changes from the whole space level to the finer sub-space level. Given the input visual and textual instances $(\boldsymbol{x}, \boldsymbol{y})$, the feature sub-space-aware pattern is defined as

$$\underset{n\in\mathcal{N}}{\mathcal{M}}(\boldsymbol{x}, \boldsymbol{y}, n), \tag{1}$$

where $\mathcal{M}(\cdot, \cdot, n)$ denotes the learning pattern with n sub-spaces. Note that the number of sub-spaces n can be various, e.g., $n \in \mathcal{N}$, meaning the different observation patterns. $\mathcal{N}$ is the set of possible values for n. Specifically, for each sub-space-aware pattern $\mathcal{M}(\boldsymbol{x}, \boldsymbol{y}, n)$, as shown in Figure 2(b1), there are three steps.

(1) Designing local sub-spaces. Formally, each sub-space is a dimensionally sub-unit divided from the entire feature representation space, and these sub-spaces are considered to imply latent semantics. For the visual representation of $\boldsymbol{x}$, its local sub-spaces can be obtained as

$$\begin{aligned} &\Gamma(\boldsymbol{x}, n) = [\hat{\boldsymbol{x}}^1, \ldots, \hat{\boldsymbol{x}}^i, \ldots, \hat{\boldsymbol{x}}^n], \\ &\text{s.t. } \textstyle\bigcup_{i=1}^{n} \hat{\boldsymbol{x}}^i = \boldsymbol{x}, \end{aligned} \tag{2}$$

where $\Gamma(\cdot, n)$ is the partitioning function that decomposes the representation $\boldsymbol{x}$ into n internal features, and $\hat{\boldsymbol{x}}^i$ denotes the i-th sub-space (the details of $\Gamma(\cdot, n)$ are in Subsection 3.1.4). Symmetrically, we can get the corresponding local sub-spaces of textual representation Y as $\Gamma(\boldsymbol{y}, n) = [\hat{\boldsymbol{y}}^1, \ldots, \hat{\boldsymbol{y}}^i, \ldots, \hat{\boldsymbol{y}}^n]$.

(2) Obtaining sub-space-aware pattern. The number of n determines how many measurable sub-spaces serve as the basic unit of semantic similarity observation. Thus, different numbers of sub-spaces constitute varying observation patterns. Based on the constructed local sub-spaces of $\boldsymbol{x}$ and $\boldsymbol{y}$, we have

$$R(\{\hat{\boldsymbol{x}}^i\}_{i=1}^n, \{\hat{\boldsymbol{y}}^i\}_{i=1}^n) = [s^1, \ldots, s^i, \ldots, s^n] \in \mathbb{R}^n, \tag{3}$$

where $R(\cdot)$ denotes the calculate the relevance s^i between the i-th sub-space of visual $\hat{\boldsymbol{x}}^i$ and textual $\hat{\boldsymbol{y}}^i$.

(3) Weighting similarity measurement. To dynamically focus on the crucial sub-spaces in the cross-modal instances, we employ a multi-layer perceptron (MLP) f to weight all observed relevances of local sub-spaces as the similarity score:

$$f(\{s^i\}_{i=1}^n) = \sum_{k=1}^{l} w_{1k}\sigma\bigg(\sum_{i=1}^{n} w_{ki}s^i\bigg), \tag{4}$$

where $\sigma(\cdot)$ is the tanh activation. $\{w_{1k}\}$ and $\{w_{ki}\}$ are the learning weight parameters.

Finally, the feature sub-space-aware pattern with n sub-spaces can be formulated as

$$\mathcal{M}(\boldsymbol{x}, \boldsymbol{y}, n) = f(R(\Gamma(\boldsymbol{x}, n), \Gamma(\boldsymbol{y}, n))). \tag{5}$$

3.1.2 *Optimal sub-space-aware pattern mining*

For various sub-space-aware patterns, to the best of our knowledge, this is the first time to reveal which kind of pattern can achieve effective image-text alignment performance gains. Yet, it is challenging since there is no prior knowledge about the relationship between observation patterns and performance gains. At the same time, exhausting all possible observation patterns is also impractical. As a feasible alternative, we propose to exploit probability statistics to indirectly decide which size level of the observation patterns has optimal performance. Specifically, the set $\mathcal{N}$ in (1) that determines the type of observation pattern is defined as hierarchical size-levels $\mathcal{N} = [n_1, \ldots, n_i, \ldots, n_m]$, $n_i = 2^i$. Note that the observation pattern at each size level can be arbitrary via the generalized partitioning function in Subsection 3.1.4. All sub-space-aware patterns constitute a parallel learning architecture.

For each pattern at the i-th size level n_i, we calculate the probability that it can achieve optimal performance as $p(n_i)$. Therefore, the optimal sub-space-aware pattern is

$$\begin{aligned} &\mathcal{M}(\boldsymbol{x}, \boldsymbol{y}, n^*), \\ &\text{s.t. } n^* \sim \underset{n_i \in \mathcal{N}}{\text{Max}}\, p(n_i), \quad i \in [1, m], \end{aligned} \tag{6}$$

where n^* denotes the size level that has the maximum probability to reach optimal performance.

Concretely, we model the probabilistic relationship between the size level of observation patterns and performance gains from a statistical perspective, as shown in Figure 2(b2). During the training phase, for each sub-space-aware pattern $\mathcal{M}(\cdot, \cdot, n_i)$, $i \in [1, m]$, its similarity measurement result in j-th epoch is regarded as a state $\boldsymbol{S}_{n_i,j}$, $j \in [1, e]$ (e is the number of training epochs). Thus, at each epoch, we pick the size level with the optimal performance gains, i.e., $n^*_{i,j} \sim \text{Max}\ \mathcal{I}(\{\boldsymbol{S}_{n_i,j}\}_{i=1}^m)$, $n^*_{i,j} \in \mathcal{N}$, where $\mathcal{I}(\cdot)$ denotes the metric (defined in (17)) to evaluate the image-text alignment accuracy that takes the measured image-text similarities as input and outputs an evaluated score (the higher the better). Then the probability of each size level is calculated as $p(n_i) = \textit{count}(n_i = n^*_{i,j})/e$, where $\textit{count}(\cdot)$ represents the number of occurrences of size level n_i in the set $\{n^*_{i,j}\}_{j=1}^e$. To make the statistics valid, we experiment with the statistical probability of n^* on different datasets, image-text alignment paradigms, and local sub-space designs, which are shown in Subsection 4.3.

3.1.3 *Optimal combined sub-space-aware patterns*

Considering that local sub-spaces at different size levels may have different measurement abilities for latent semantics, e.g., from low-level to high-level, there are likely complementary relationships between various sub-space-aware patterns. In this section, we are devoted to solving which sub-space-aware patterns combinatorially contribute to the optimal image-text alignment performance.

Based on the sub-space-aware patterns $\mathcal{M}(\cdot, \cdot, n_i)$, $i \in [1, m]$, we can obtain the training result of each pattern as $\boldsymbol{S}_{n_i,j}$, $j \in [1, e]$, which is regarded as an optional state for the combined optimal performance. Our goal is to dynamically mine and search for a set of levels to achieve optimal combined sub-space-aware learning. Consequently, we formulate it as a discrete dynamic searching problem:

$$\begin{aligned} &\underset{\kappa_i}{\text{Max}}\ \ \mathcal{I}\Big(\sum_{i=1}^{m} \kappa_i \cdot \boldsymbol{S}_{n_i,j}\Big), \\ &\text{s.t.}\ \ \sum_{i=1}^{m} \kappa_i \leqslant m, \kappa_i \in \{0, 1\}, \end{aligned} \tag{7}$$

where κ_i is the discrete decision variable, i.e., $\kappa_i \in \{0, 1\}$. For the size level n_i of a sub-space-aware pattern, it is selected when the corresponding κ_i is 1, otherwise it is discarded; m is the number of total patterns; as defined in Subsection 3.1.2, $\mathcal{I}(\cdot)$ is the metric (defined in (17)) to evaluate the alignment accuracy of measured image-text similarities $\boldsymbol{S}_{n_i,j}$ (the higher the better).

We solve the above problem by deriving the optimal state transition equation as

$$\begin{aligned} &\boldsymbol{S}^*_k = \text{Max}\{\mathcal{I}(\boldsymbol{S}^*_{k-1}), \mathcal{I}(\boldsymbol{S}^*_{k-1} + \boldsymbol{S}_k)\}, \quad k \in [1, m], \\ &\text{s.t.}\quad \{\boldsymbol{S}_k\}_{k=1}^m = \text{Sort}\Big\{\Big\{\underset{j}{\text{Max}}(\mathcal{I}(\boldsymbol{S}_{n_i,j}))\Big\}_{i=1}^m\Big\}, \end{aligned} \tag{8}$$

where $\boldsymbol{S}^*_m$ is the final optimal state, $\boldsymbol{S}^*_0 = 0$, and $\text{Sort}(\cdot)$ denotes a descending sort operation. For S^*_m, during the k-step state transition process, we obtain the corresponding κ_i value according to whether state $\boldsymbol{S}_{n_i,j}$ (belonging to size level n_i) has contributed. That is

$$\kappa_i = \begin{cases} 1, & \text{when } \mathcal{I}(\boldsymbol{S}^*_{k-1} + \{\boldsymbol{S}_{n_i,j}\}_k) > \mathcal{I}(\boldsymbol{S}^*_{k-1}), \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Finally, as shown in Figure 2(b3), according to the search result $\{\kappa_i\}_{i=1}^m$, we determine the corresponding size level set $\mathcal{N}^* = \{n_i \cdot \kappa_i\}$ that models the optimal complementarity of different sub-space-aware patterns for the cross-modal similarity measurement:

$$\underset{n \in \mathcal{N}}{\mathcal{M}}(\boldsymbol{x}, \boldsymbol{y}, n) = \sum \mathcal{M}(\boldsymbol{x}, \boldsymbol{y}, n^*_i), \quad n^*_i \in \mathcal{N}^*. \tag{10}$$

In addition, we also provide a version with continuous weight learning to obtain the optimal combination patterns. Compared with it, our proposed discrete dynamic search method can produce a more concise and sparse size level set, which has better performance as verified in Subsection 4.3.

3.1.4 *Partitioning function*

In this subsection, we focus on the specific implementation of the partitioning function $\Gamma(\cdot, n)$. For the input visual/textual feature representation, it aims to decompose it into n internal feature sub-spaces. Here, we give two methods: (1) generalized partitioning and (2) average partitioning. The first method obtains feature sub-space-aware patterns with arbitrary internal features by a random generator, while the second method obtains a relatively simpler average internal feature.

First, given the number n, we resort to the random probability generator to obtain $[p_1, \ldots, p_i, \ldots, p_n]$, which satisfy $p_i \in [0, 1]$ and $\sum_{i=1}^{n} p_i = 1$. Therefore, for each local sub-space $\hat{\boldsymbol{x}}^i$ in (2), it is defined as $\hat{\boldsymbol{x}}^i = \boldsymbol{x}_{l_{i-1}:l_i}, l_i = \lfloor (\sum_{j=0}^{i} p_j) \cdot d \rfloor$, where $[l_{i-1} : l_i]$ represents the slice interval in the entire d-dimensions of X, and $\lfloor \cdot \rfloor$ means rounding down. Note that the randomly generated p_i is likely to be 0, so this random partitioning can generalize the number of local sub-spaces around n. Second, as a special case of the first one, when all p_i are equal, it is the average decomposition.

3.2 Generalization on existing paradigms

Note that our proposed OPEN framework enjoys the merit of plug-and-play, which means it is easily applied to existing image-text alignment approaches. We next describe how to equip the OPEN framework $\mathcal{M}(\cdot, \cdot, n)$, $n \in \mathcal{N}$ to the two most typical paradigms for image-text alignment, i.e., O2O and M2M, depicted in Figures 2(c1) and (c2), respectively. The entire training process of our framework is summarized in Algorithm 1.

Algorithm 1 Training process of our proposed OPEN framework.

Input: Images data $\{\boldsymbol{I}\}$ and texts data $\{\boldsymbol{T}\}$.
Output: Neural network parameters, $\{\kappa_i\}_{i=1}^{m}$ and $\{p(n_i)\}_{i=1}^{m}$.
Start training with setting total training epochs e and total patterns m;
for $j = 1, \ldots, e$ **do**
  Construct image-text pairs $\boldsymbol{I}$-$\boldsymbol{T}$;
  Extract visual and textual features $(\boldsymbol{x}, \boldsymbol{y})$;
  Construct sub-space-aware pattern $\mathcal{M}(\boldsymbol{x}, \boldsymbol{y}, n)$, $n \in \mathcal{N}$, $\mathcal{N} = \{n_i\}$, $i \in [1, m]$;
  Obtain the local sub-space via (2);
  Obtain the sub-space-aware pattern via (3);
  Weighted similarity measurement via (4);
  Obtain image-text similarity via O2O in (11) or M2M in (15);
  Calculate the loss backward via (16);
  Record the evaluated similarity results as the state $\boldsymbol{S}_{n_i,j}$, $i \in [1, m]$, $j \in [1, e]$;
end for
Calculate the statistical probabilities $\{p(n_i)\}_{i=1}^{m}$ of each pattern to be optimal via (6);
Calculate the discrete variable $\{\kappa_i\}_{i=1}^{m}$ to determine optimal combined patterns via (7);
End training.

3.2.1 *One-to-one*

The O2O paradigm tends to represent images or texts as holistic features that measure cross-modal similarity to learn global-level semantic alignment. Specifically, for the feature representations of a given image-text pair, the input image and text are represented as $\boldsymbol{I} \in \mathbb{R}^{1\times d}$ and $\boldsymbol{T} \in \mathbb{R}^{1\times d}$, where d is the dimension of feature vectors. Then, the image-text similarity can be directly measured as

$$S(\boldsymbol{I}, \boldsymbol{T}, n) = \underset{n\in\mathcal{N}}{\mathcal{M}}(\boldsymbol{I}, \boldsymbol{T}, n), \tag{11}$$

where $\mathcal{M}(\cdot, \cdot, n)$ denotes the proposed feature sub-space-aware pattern.

3.2.2 *Many-to-many*

The M2M paradigm focuses on discovering semantic alignments between all image regions (patches) and text words at the fragment level, and the overall image-text similarity is inferred from the similarity score of each fragment. Generally, most M2M methods are based on the attention mechanism.

Specifically, for the feature representations of a given image-text pair, which consists of p salient image regions (patches) and q text words, the image is represented by a set of visual features $\boldsymbol{I} = \{\boldsymbol{v}_i \mid i \in [1,p], \boldsymbol{v}_i \in \mathbb{R}^d\}$, while the text is represented by a set of textual features $\boldsymbol{T} = \{\boldsymbol{t}_j \mid j \in [1,q], \boldsymbol{t}_j \in \mathbb{R}^d\}$, where d is the dimension of feature vectors. First, similarities of all fragment pairs are computed as s_{ij}, $i \in [1,p]$, $j \in [1,q]$. Then, there are two directions to measure the image-text similarity, i.e., image-to-text (i2t) and text-to-image (t2i).

For image-to-text that uses each visual fragment $\boldsymbol{v}_i$ as a query, the M2M paradigm aims to weight the relevant word features as

$$\hat{\boldsymbol{t}}_i = \sum_{j=1}^{q} w_{ij}\boldsymbol{t}_j, \quad \text{s.t. } w_{ij} = \frac{\exp(\lambda \hat{s}_{ij})}{\sum_{j=1}^{q}\exp(\lambda \hat{s}_{ij})}, \tag{12}$$

where w_{ij} denotes the attention weight between the i-th region and the j-th word, $\hat{s}_{ij} = [s_{ij}]_+ / \sqrt{\sum_{j=1}^{q}[s_{ij}]_+^2}$, $[\cdot]_+ = \max(\cdot, 0)$, and λ is a scaling parameter. Thus, the similarity score of i-th visual fragment can be measured by $\boldsymbol{v}_i$ and the weighted text feature $\hat{\boldsymbol{t}}_i$ as

$$s_i^{\text{i2t}} = \underset{n\in\mathcal{N}}{\mathcal{M}}^{\text{i2t}}(\boldsymbol{v}_i, \hat{\boldsymbol{t}}_i, n), \tag{13}$$

where note that it is symmetric for each word as the query to measure its similarity score, i.e., $s_j^{\text{t2i}} = \mathcal{M}_{n\in\mathcal{N}}^{\text{t2i}}(\boldsymbol{t}_j, \hat{\boldsymbol{v}}_j, n)$.

Since there are two directions of sub-space-aware patterns to measure similarity between vision and language, the word-region similarities are calculated by

$$s_{ij} = \underset{n\in\mathcal{N}}{\mathcal{M}}^{\text{i2t}}(\boldsymbol{v}_i, \boldsymbol{t}_j, n) + \underset{n\in\mathcal{N}}{\mathcal{M}}^{\text{t2i}}(\boldsymbol{v}_i, \boldsymbol{t}_j, n), \quad i \in [1,p], \quad j \in [1,q]. \tag{14}$$

Typically, the overall image-text similarity is composed of similarity scores of all regions and words as

$$S(\boldsymbol{I}, \boldsymbol{T}, n) = \frac{1}{p}\sum_{i=1}^{p} s_i^{\text{i2t}} + \frac{1}{q}\sum_{j=1}^{q} s_j^{\text{t2i}}. \tag{15}$$

3.3 Objective function

Following most existing image-text alignment methods, we adopt the hardest hinge-based triplet ranking loss as the objective function, which aims to force the similarities of positive (aligned) image-text pairs to be higher than those of any other negative (misaligned) pairs by a fixed margin. Given a positive image-text pair $(\boldsymbol{I}, \boldsymbol{T})$ and all its negative pairs $(\boldsymbol{I}, \boldsymbol{T}^-)$ and $(\boldsymbol{I}^-, \boldsymbol{T})$, the hardest samples are selected by $\boldsymbol{I}_h^- = \arg\max_{p\neq\boldsymbol{T}} S(\boldsymbol{I}, p, n)$ and $\boldsymbol{T}_h^- = \arg\max_{q\neq\boldsymbol{I}} S(q, \boldsymbol{T}, n)$. Thus, the objective function is written as

$$L(\boldsymbol{I}, \boldsymbol{T}, n) = [\gamma - S(\boldsymbol{I}, \boldsymbol{T}, n) + S(\boldsymbol{I}, \boldsymbol{T}_h^-, n)]_+ + [\gamma - S(\boldsymbol{I}, \boldsymbol{T}, n) + S(\boldsymbol{I}_h^-, \boldsymbol{T}, n)]_+, \tag{16}$$

where γ denotes the margin parameter, $[\cdot]_+ = \max(\cdot, 0)$, and $S(\boldsymbol{I}, \boldsymbol{T}, n)$ is the similarity measured by the sub-space-aware pattern $\mathcal{M}_{n\in\mathcal{N}}(\cdot, \cdot, n)$.

3.4 Feature extraction

For a fair comparison, we adopt the three most used feature extraction settings in existing methods [9, 13, 14, 22, 26–32, 44–48]. They are listed as follows.

(1) **BiGRU+BUTD.** BUTD refers to the bottom-up attention [49]. The faster R-CNN model in conjunction with ResNet-101, which is pre-trained on visual genome [50], is adopted to detect objects and other salient regions. Top-K ($K = 36$) local regions are selected with respect to each image, and we obtain their mean-pooled convolutional features with 2048 dimensions. Then, a fully connected layer is used to transform each region into a final d-dimensional feature. BiGRU refers to the common textual encoder, where each word is first represented as a one-hot encoding and embedded into an embedding vector; then the vectors are fed into a bi-directional gated recurrent unit to integrate the forward and backward contextual information. The final word representation is the average of the bi-directional hidden states. In addition, following some existing studies [13, 14], we also report the performance of the BiGRU encoder enhanced by pre-trained GloVe vectors [51].

(2) **BERT+BUTD.** The visual encoder uses the same BUTD as above, while the textual encoder uses more advanced BERT [52], which is a transformer-based model pre-trained on large-scale Wikipedia and Bookcorpus. We add a fully connected layer for the last layer of pre-trained BERT to obtain word features.

(3) **CLIP-based BERT+ViT.** Different from the BUTD that requires explicit object detection, the visual encoder exploited vision Transformer (ViT) [53] directly uses the image patches as input. Moreover, to discover

the fine-grained vision-language relationships, we use all output tokens from the CLIP-based ViT encoder and the CLIP-based BERT encoder as visual and textual features, respectively. The version of CLIP is selected as ViT-B/16 for fine-tuning [4].

3.5 Discussion

Here, we comprehensively discuss the advantages of the proposed OPEN framework and give an in-depth theoretical analysis. In addition, we also analyze the potential limitations of OPEN and future research in combination with large language models (LLMs).

(1) **Novelty and advantages**. Different from existing methods that typically leverage cosine or Euclidean distance to measure similarity between visual and textual features [6–8, 27, 29, 34, 35, 43], where the modality feature is directly examined in the whole representation space, and motivated by the discovery that partial local dimensions, forming sub-spaces with the potential semantic representation tendency, contain more important semantic measurement information, we propose a novel optimal feature sub-space-aware similarity learning framework, which takes a forward step to focus on the sub-space composed of local dimensions within the modality representation. We highlight our advantages as follows. (i) Changing the cross-modal semantic similarity measurement from the prior whole representation space to the sub-space composed of partial local dimensions enables a more subtle semantic alignment similarity measurement. Moreover, we reveal that the pattern at the middle size-level has the maximum probability of being optimal, which can be directly used as a prior to improve semantic alignment. (ii) An optimal combination pattern mining method is designed, which dynamically discovers the optimal complementarities of different size-level sub-space-aware patterns, thereby producing more comprehensive and accurate similarity measurements. (iii) Our proposed OPEN framework has the merit of plug-and-play, and we devise the way to equip OPEN to the existing two typical image-text matching paradigms, verifying its effectiveness.

(2) **In-depth theoretical analysis**. In the representation space spanned by d-dimensions, from the perspective of the dimensional space involved in semantic similarity examination, in contrast to existing similarities, e.g., cosine or Euclidean distance, which only leverage a single level of granularity, i.e., the whole space, our proposed method introduces new similarity examination granularities, namely, the sub-space composed of dimensions of different size levels. In theory, the proposed framework increases the examination granularity of semantic alignment from the entire space to the different sub-spaces composed of internal local dimensions, thereby improving semantic alignment.

(3) **Potential limitations**. We analyze two potential limitations of the proposed method in real-world applications. (i) Image-text data with complex semantic polysemy: for polysemous words with multiple semantics in the real world, e.g., 'mouse' can mean both a mouse and a rat, the improvement of visual-linguistic cross-modal similarity calculation may not be sufficient to model this semantic ambiguity. (ii) Image-text data with different aesthetic qualities: for data with different qualities, especially in terms of aesthetics, i.e., whether the image has a suitable composition or a high-level visual communication effect, the proposed method may not have the ability to distinguish. However, the proposed method can be enhanced by combining multi-view modeling [54, 55] and aesthetic image annotation [56].

(4) **Future research in combination with LLMs**. First, LLMs can be leveraged to generate more precise textual semantic embeddings, thereby enhancing image-text alignment. Additionally, from the perspective of semantics and sub-spaces that focus on their semantic similarity, consider the hierarchical relationship between different semantics (e.g., 'person' can represent both sub-objects 'man' and 'woman'), which can be obtained through LLMs with extensive prior knowledge of sub-object relationships. Based on this semantic sub-object relationship, how to explore the relationship between their hierarchical semantically important sub-spaces, such as the sub-space that focuses on the semantics of 'person' and the sub-space that focuses on the semantics of sub-objects like 'man' or 'woman'. Combining the explicit semantic hierarchical relationship parsed by LLMs to constrain the semantically important sub-spaces may further enhance the performance, which can be explored in future work.

4 Experiments

4.1 Datasets and implementation details

4.1.1 *Datasets*

Two generic datasets are used to validate the effectiveness of our proposed OPEN, where each dataset consists of a large number of images, and every image is annotated with five ground-truth descriptions. (1) Flickr30K: there are 31000 images and 155000 texts. We follow the standard split protocol in [5], using 1000 images for testing,

1000 images for validation, and 29000 images for training. (2) MS-COCO: it contains 123287 images and 616435 sentences. Based on previous work [57], we split it into 5000 test images, 5000 validation images, and 113287 training images. Performance on MS-COCO is calculated by averaging over 5-fold of 5k test images.

4.1.2 *Evaluation metric*

The commonly used evaluation metric in image-text matching is R@K ($K = 1, 5, 10$), which means the percentage of ground truth in the retrieved top-K lists. Moreover, following existing studies [3, 6, 24], we also adopt rSum that is the sum of all R@K in both text retrieval, i.e., image-to-text, and image retrieval, i.e., text-to-image, directions as

$$\text{rSum} = \underbrace{\text{R@1} + \text{R@5} + \text{R@10}}_{\text{image-to-text}} + \underbrace{\text{R@1} + \text{R@5} + \text{R@10}}_{\text{text-to-image}}, \tag{17}$$

which reflects the overall image-text matching performance.

4.1.3 *Implementation details*

The proposed OPEN is implemented using PyTorch. All experiments are conducted on an NVIDIA GeForce RTX A40 GPU and adopt the Adam optimizer with 0.0005 as the initial learning rate. Without additional explanation, the proposed OPEN is built based on the O2O method VSE∞ [3]. In addition, we also verified the effectiveness of our method on other M2M methods; see Subsection 4.3. For O2O-based OPEN, the training epoch is set to 20 with the learning rate decaying by 10% every 10 epochs, and the batch size is set to 128. For M2M-based OPEN on both Flickr30K and MS-COCO, the training epoch is set to 20 with the learning rate decaying by 10% every 10 epochs; the batch size is set to 32; the scaling parameter λ is set to 20. For fair comparisons, the dimension d of visual or textual features is set to the same as in existing methods. The i-th size-level is determined by $n_i = 2^i$, $i \in [1, m]$. m is selected by $2^{m+1} = d$, i.e., $m = 9$. The lengths of the two full connection layers are n_i and $l = \lceil n_i/2 \rceil$ ($\lceil \cdot \rceil$ means rounding up). The margin parameter γ is set as 0.2.

In all feature extraction settings, including (1) BiGRU+BUTD, (2) BERT+BUTD, and (3) CLIP-based BERT+ViT, except for the bottom-up attention BUTD, using pre-trained visual object detector Faster-RCNN and visual encoder ResNet-101 to extract image salient region features (exactly following the existing methods [3]), which can be regarded as frozen, the rest are not frozen. As in the existing methods, in all settings, a mapping layer is added after the visual encoder and text encoder, respectively, to map the visual and text features to the common representation space.

4.2 Quantitative results

4.2.1 *Comparison with state-of-the-art methods*

To adequately verify the effectiveness of our proposed OPEN framework, we report the performance comparisons with the recent state-of-the-art models on the two benchmarks, where we provide two versions of single models OPEN(g-) and OPEN(a-), with respect to different partitioning methods of local sub-spaces, namely generalized and average partitioning defined in Subsection 3.1.4. Existing methods are divided into four types based on their feature backbones for fair comparisons.

Table 1 shows the quantitative results of our method on the Flickr30K test set and MS-COCO 1k test set. We can observe that the proposed OPEN significantly outperforms state-of-the-art models [9, 13, 14, 22, 26–32, 44–48] with respect to all the evaluation metrics on both datasets. Compared with the existing state-of-the-art methods in different feature backbone settings, i.e., BiGRU+BUTD, BiGRU(GloVe)+BUTD, and BERT+BUTD, our approaches yield relative average improvements of 3.78% and 1.78% in rSum on Flickr30K and MS-COCO, respectively. Compared with the baseline method VSE∞ on the two datasets, our framework yields relative average improvements of 5.35% and 8.25% in R@1 on image-to-text and text-to-image directions, respectively. As shown in Table 2, OPEN outperforms state-of-the-art models [9, 13, 14, 22, 26–31, 44–46] on all evaluation metrics in testing on the MS-COCO 5k test set. The consistent improvements of OPEN demonstrate its effectiveness and superiority.

Besides, based on the vision-language pre-trained feature backbone, as shown in Table 3, we verify the effectiveness of the proposed method OPEN based on two settings, CLIP-ViT-base-224+CLIP-BERT-base, and CLIP-ViT-large-384+CLIP-BERT-large [4]. From the results, we can see that compared with the baseline methods [4, 35–40], where the modality feature is directly examined in the whole representation space via cosine or Euclidean distance, our proposed feature sub-space-aware visual-language cross-modal similarity learning framework can achieve better alignment performance.

Table 1 Comparisons with state-of-the-art methods on the Flickr30K test set and the MS-COCO 1k test set. $*$ means ensemble results. The best results are in bold.

Method	Flickr30K dataset							MS-COCO (1k) dataset						
	Image-to-text			Text-to-image			rSum	Image-to-text			Text-to-image			rSum
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
BiGRU+BUTD														
SGRAF*('21) [9]	77.8	94.1	97.4	58.5	83.0	88.8	499.6	79.6	96.2	98.5	63.2	90.7	96.1	524.3
CMCAN('22) [44]	77.5	94.3	96.9	58.8	82.9	88.9	499.3	79.9	96.6	98.8	63.3	90.4	96.2	525.2
NAAF('22) [13]	75.9	93.6	97.7	55.5	81.0	87.9	491.5	76.8	95.2	98.2	61.3	90.6	96.0	518.2
DRCE*('23) [45]	77.3	94.6	97.4	58.5	83.1	89.3	500.2	79.1	96.4	**99.0**	63.6	90.3	95.9	524.3
ESL('24) [28]	**78.5**	95.4	97.8	56.6	83.1	89.4	500.8	78.8	96.3	98.8	62.4	90.1	95.7	522.1
OPEN(-a)	78.4	95.0	**97.9**	**58.8**	**84.4**	**90.9**	**505.4**	**80.2**	**96.5**	98.7	**64.1**	**90.9**	**96.2**	**526.6**
OPEN(-g)	79.1	**95.7**	**98.1**	**58.9**	**85.1**	**90.6**	**507.8**	**80.1**	**96.7**	98.9	**63.9**	**90.7**	**95.9**	**526.1**
BiGRU(GloVe)+BUTD														
HREM('23) [46]	79.5	94.3	97.4	59.3	85.1	91.2	506.8	80.0	96.0	98.7	62.7	90.1	95.4	522.8
CHAN('23) [47]	79.7	94.5	97.3	60.2	85.3	90.7	507.8	79.7	96.7	98.7	63.8	90.4	95.8	525.0
DSRLN('24) [26]	78.8	95.3	97.7	59.9	85.8	91.9	509.5	79.0	96.3	98.7	64.0	90.6	96.0	524.6
LMG-C('24) [32]	79.6	95.5	98.5	61.0	84.5	90.5	509.6	78.2	96.3	98.6	63.1	89.9	95.5	521.6
ESL('24) [28]	82.0	95.5	98.5	60.6	86.3	91.7	514.7	79.2	96.6	98.7	63.3	90.4	96.0	524.2
NUIF-d('24) [30]	81.8	95.7	98.0	59.0	83.9	89.9	508.3	79.9	96.7	99.0	63.9	90.4	95.8	525.7
OPEN(-a)	**82.3**	**95.9**	**98.6**	60.9	**86.4**	91.5	**515.6**	79.9	**96.8**	**99.1**	**64.4**	**91.2**	**96.3**	**527.7**
OPEN(-g)	**83.0**	**95.8**	**98.9**	**61.1**	**87.3**	**91.9**	**518.0**	**80.1**	96.6	98.7	**64.5**	**91.5**	**96.0**	**527.4**
BERT+BUTD														
VSRN++*('22) [48]	79.2	94.6	97.5	60.6	85.6	91.4	508.9	77.9	96.0	98.5	64.1	91.0	96.1	523.6
CHAN('23) [47]	80.6	96.1	97.8	63.9	87.5	92.6	518.5	81.4	96.9	98.9	66.5	92.1	96.7	532.6
HREM('23) [46]	83.3	96.0	98.1	63.5	87.1	92.4	520.4	81.1	96.6	98.9	66.1	91.6	96.5	530.7
X-Dim('23) [14]	83.6	96.0	98.4	65.0	88.7	93.1	524.8	82.2	97.2	99.1	66.9	92.0	96.6	534.0
Multi('25) [29]	83.2	96.5	98.5	63.6	87.9	92.7	522.4	80.6	96.7	98.8	66.1	91.5	96.3	530.0
USER('24) [22]	82.7	97.0	98.3	63.1	86.7	92.1	519.9	82.8	96.8	98.8	66.1	90.6	95.6	530.5
HiCer('24) [27]	84.0	96.6	98.4	64.4	88.1	93.0	524.6	82.6	96.9	98.8	67.9	91.5	96.2	533.8
ESL('24) [28]	83.5	96.3	98.4	65.1	87.6	92.7	523.7	82.2	97.1	99.0	66.2	91.9	96.7	533.1
IMEB('24) [31]	84.2	96.7	98.4	64.0	88.0	92.8	524.1	82.4	96.9	99.0	66.7	91.9	96.6	533.5
NUIF-d('24) [30]	83.9	96.5	98.2	67.9	89.2	93.6	529.4	83.3	97.3	98.9	69.2	92.7	96.9	538.2
OPEN(-a)	**84.3**	**97.2**	**99.5**	**67.9**	**89.3**	**94.9**	**532.2**	**84.2**	**97.3**	98.9	**69.7**	**93.0**	**97.5**	**540.6**
OPEN(-g)	**83.9**	**97.7**	**99.7**	**67.9**	**89.9**	**94.4**	**533.5**	**84.0**	**97.4**	**99.2**	**69.9**	**92.7**	**97.3**	**540.5**

Moreover, we also compare with state-of-the-art visual-language pre-training models [36–40] on the Flickr30K test dataset. We can see that OPEN brings competitive performances compared to the mainstream VLP models. Our OPEN achieves R@1 = 96.1% for image-to-text retrieval and R@1 = 85.7% for text-to-image retrieval, which brings good values for real applications.

The performance improvements on all experiment settings, i.e., different partitioning methods of local bases, baselines of image-text paradigms, and datasets, show the superiority of our general framework for non-uniformly measuring the similarity between visual and textual features, which makes the semantic alignment measure take a forward step to finer local bases within textual/visual features, enabling more accurate similarity measurement.

4.2.2 *Optimality of sub-space-aware pattern*

As we analyzed in Subsection 3.1.1, the different size levels of sub-space enable the similarity measurement of cross-modal semantic alignment with varying observation patterns. Thus, we expect to mine and reveal which kind of sub-space-aware learning pattern can bring the optimal performance in Subsection 3.1.2.

To comprehensively and clearly reveal the relationship between different size-level patterns and performance, we show the performance states of all patterns at each epoch in Figure 3 under different baseline models. Experimentally, we can observe that the middle size-level, i.e., $n_i, i \in \{4, 5, 6\}$, can usually be the peak of the state surface in Figure 3. At the same time, we statistically reveal the probability that each pattern can achieve optimal performance, as shown in Figure 3. From the statistical perspective, the pattern at the middle size-level has the maximum probability to be optimal.

To reveal the optimal pattern more accurately, we further calculate the probabilistic relationship on different

Table 2 Comparisons on MS-COCO 5k test set. $*$ means ensemble results. The best results are in bold.

Method	Image-to-text			Text-to-image			rSum
	R@1	R@5	R@10	R@1	R@5	R@10	
BiGRU+BUTD							
SGRAF('21) [9]	57.8	–	91.6	41.9	–	81.3	272.6
CMCAN('22) [44]	57.6	84.6	91.1	41.0	70.1	81.2	425.6
NAAF('22) [13]	57.3	84.3	91.9	41.6	70.5	81.7	427.3
DRCE*('23) [45]	56.2	83.0	90.9	40.3	69.5	80.6	420.5
ESL('24) [28]	58.0	84.6	91.8	40.7	70.1	80.8	426.0
OPEN(-g)	**59.4**	**85.3**	**92.1**	**42.0**	**71.7**	**81.9**	**432.3**
OPEN(-a)	**59.6**	**85.5**	**92.1**	41.8	**71.8**	**82.2**	**432.9**
BiGRU(GloVe)+BUTD							
NAAF ('22) [13]	55.8	84.0	91.5	41.4	70.2	80.3	423.3
HREM ('23) [46]	58.9	85.3	92.1	40.0	70.6	81.2	428.1
DSRLN('24) [26]	57.7	84.9	**92.4**	41.2	71.7	82.3	430.2
ESL ('24) [28]	58.6	85.4	92.0	41.6	70.8	81.3	429.8
NUIF-d('24) [30]	59.3	85.5	92.0	41.9	71.3	81.8	431.8
OPEN(-g)	**60.0**	84.6	92.1	**42.5**	**72.6**	**82.9**	**434.7**
OPEN(-a)	**59.7**	**85.8**	92.3	**42.2**	**72.3**	**82.6**	**434.9**
BERT+BUTD							
X-Dim('23) [14]	62.7	87.0	93.3	45.1	73.6	83.4	445.1
HREM('23) [46]	64.0	88.5	93.7	45.4	75.1	84.3	450.9
Multi('25) [29]	59.8	86.0	92.8	43.4	73.6	83.6	439.2
USER('24) [22]	63.7	87.4	93.5	44.8	73.4	82.7	445.5
HiCer('24) [27]	63.2	87.6	93.8	46.8	75.5	84.2	451.1
ESL ('24) [28]	63.0	87.4	93.5	44.3	74.1	84.0	446.3
IMEB('24) [31]	62.8	87.8	93.5	44.9	74.6	84.0	447.6
NUIF-d('24) [30]	65.2	88.8	94.2	48.3	76.8	85.7	459.1
OPEN(-g)	65.0	88.7	**94.9**	**49.7**	76.4	85.5	**460.2**
OPEN(-a)	**65.8**	**88.8**	**94.7**	**49.4**	**76.8**	**85.9**	**461.4**

Table 3 Comparisons with the pre-trained visual-language models (VLM). # represents the zero-shot. 'ft' denotes the fine-tuned baseline. Equipped with OPEN, and bold indicates performance that exceeds the fine-tuned baseline. The results of proposed method are in bold.

Method	Image-to-text			Text-to-image		
	R@1	R@5	R@10	R@1	R@5	R@10
UNITER [36]	87.3	98.0	99.2	75.6	94.1	96.8
VILLA [37]	87.9	97.5	98.8	76.3	94.2	96.8
UNIMO [38]	89.4	98.9	99.8	78.0	94.2	97.1
SGG-MVAR [35]	93.9	99.3	99.8	82.4	96.4	98.4
ALIGN [39]	95.3	99.8	100.0	84.9	97.4	98.6
ALBEF [40]	95.9	99.8	100.0	85.6	97.5	98.9
CLIP-ViT-base-224+CLIP-BERT-base						
$\text{CLIP}_{\#}$ [4]	81.3	96.4	98.5	62.2	85.7	91.7
Baseline_{ft}	91.7	99.0	99.5	79.1	95.2	97.6
$+\text{OPEN(-a)}_{ft}$	**92.1**	**99.2**	**99.8**	**80.5**	**95.5**	**97.9**
$+\text{OPEN(-g)}_{ft}$	**92.0**	**99.4**	**99.9**	**80.3**	**95.8**	**97.8**
CLIP-ViT-large-384+CLIP-BERT-large						
$\text{CLIP}_{\#}$ [4]	86.7	98.0	99.1	67.0	88.8	93.3
Baseline_{ft}	94.5	99.1	99.5	84.5	97.0	98.1
$+\text{OPEN(-a)}_{ft}$	**95.9**	**99.7**	**99.9**	**85.8**	**97.5**	**98.9**
$+\text{OPEN(-g)}_{ft}$	**96.1**	**99.8**	**100.0**	**85.7**	**97.8**	**99.0**

datasets, i.e., Flickr30K and MS-COCO, which are depicted in Figures 4(a) and (b), respectively. The overall probabilistic relationship of cross both datasets is shown in Figure 4(c). We can get that all of them have a similar statistical significance with respect to the pattern at the middle size-level, where n_i, $i \in \{5, 6\}$ are more important than n_i, $i = 4$. The reason is that too few sub-spaces cannot exhibit the power of finer semantic

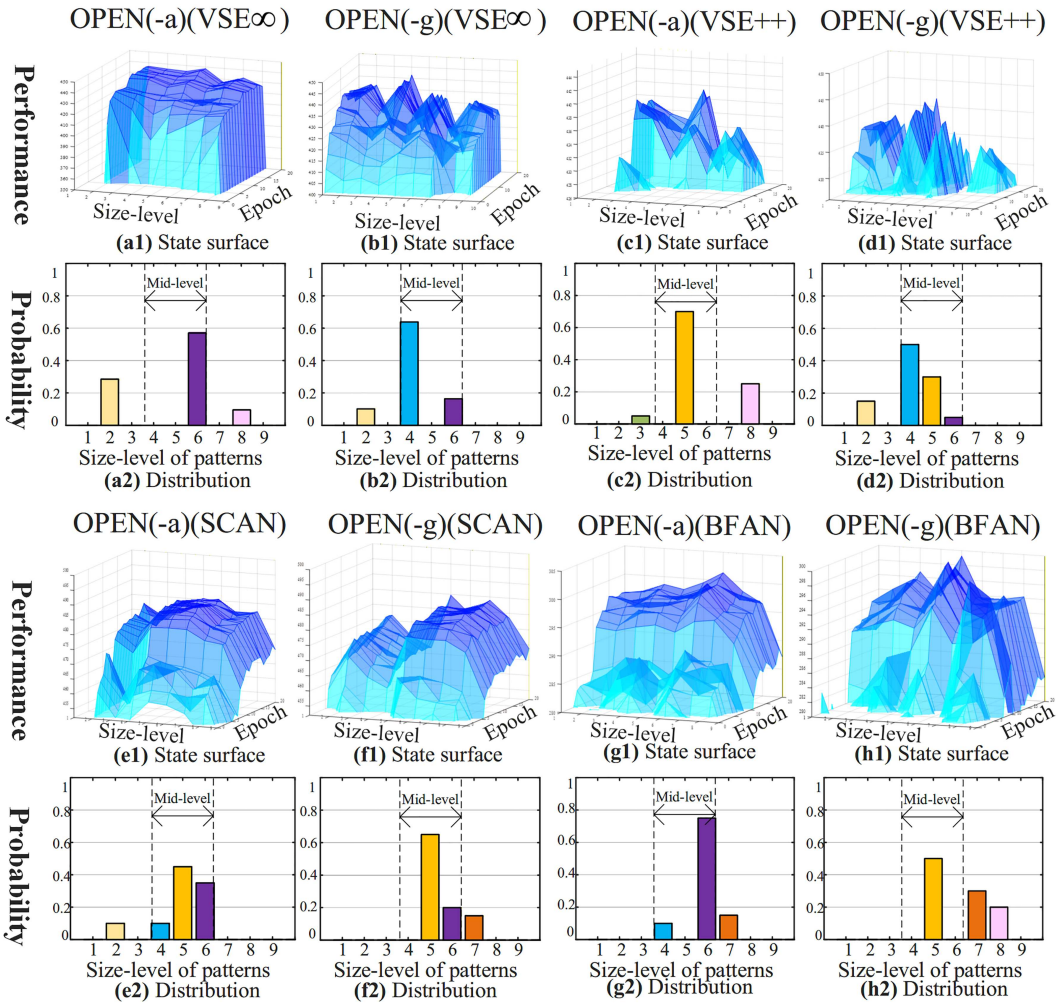

Figure 3 (Color online) Relationship between different size-level patterns and performance on various baselines. (*-1) depicts the state surface of all patterns; (*-2) models the statistical probability that each pattern can be optimal.

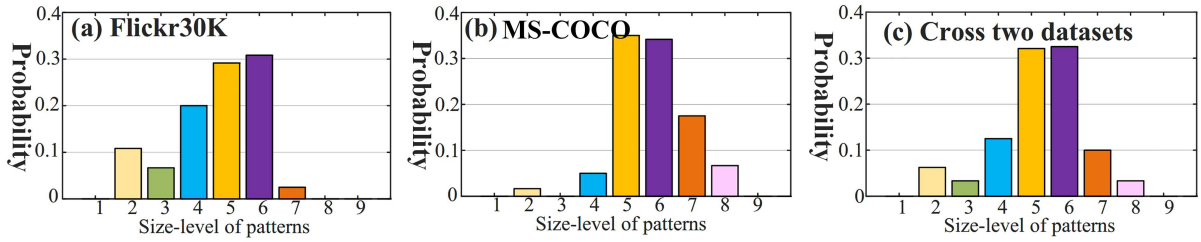

Figure 4 (Color online) Statistical probability across different datasets.

measurement, while too many sub-spaces lead to the degradation of their representation ability. This reveals a prior for similarity observation patterns in internal sub-spaces of cross-modal features. In Subsection 4.3.1, we also verify the effectiveness of directly utilizing this prior.

4.3 Ablation study

4.3.1 *Effect of framework designing*

As shown in Table 4, both (1) average and (2) generalized OPEN models can bring considerable performance improvements, demonstrating the robustness of our framework. The following models are investigated based on the average version of our OPEN framework with BiGRU+BUTD backbone.

(3) The hierarchical size-levels of sub-space designing. When the hierarchical structure is not complete, e.g.,

Table 4 Effect study of framework designing on Flickr30K and MS-COCO.

Method	Flickr30K dataset				MS-COCO dataset			
	Image-to-text		Text-to-image		Image-to-text		Text-to-image	
	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10
Baseline	76.5	94.2	56.4	83.4	78.5	98.7	61.7	95.6
(1) OPEN (average)	78.4	97.9	58.8	90.9	80.2	98.7	64.1	96.2
(2) OPEN (generalized)	79.1	98.1	58.9	90.6	80.1	98.9	63.9	95.9
(3) OPEN (w/o hierarch)	76.1	93.7	56.3	84.5	78.4	94.6	62.2	96.5
(4) OPEN (FC-learning)	68.9	95.4	53.2	86.0	73.4	90.2	49.0	89.6
(5) OPEN (w/o mining)	77.4	97.0	57.5	89.9	79.1	97.2	63.3	95.4
(6) OPEN (w-mid-levels)	76.9	95.0	57.0	86.5	78.8	96.1	62.9	95.7
(7) OPEN (CW-learning)	77.9	95.4	58.2	89.4	79.6	97.7	63.5	95.9

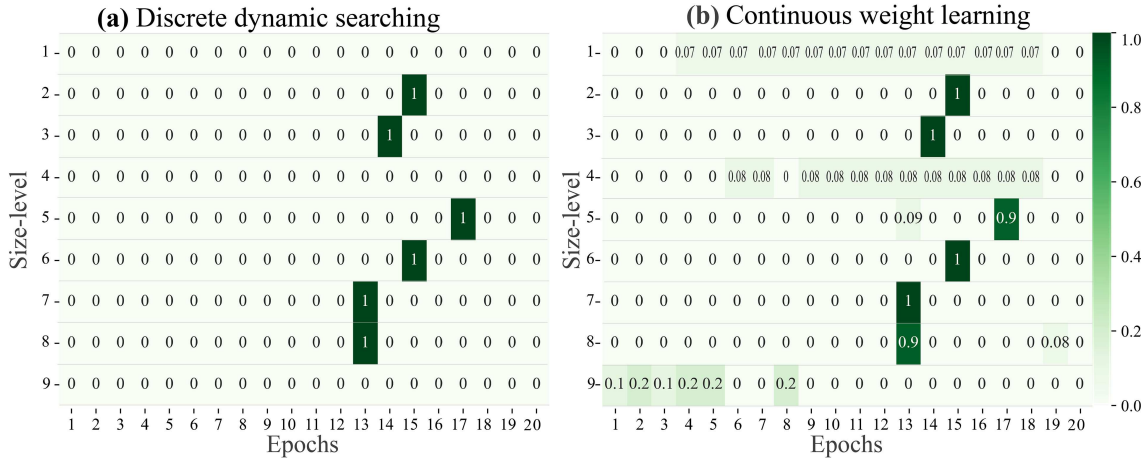

Figure 5 (Color online) Visualization of the size-level set mining for optimal learning, where (a) is the discrete results of our proposed OPEN, and (b) is a continuous weight learning version.

$m = 3$, relative to the full hierarchies, we can find performance degradation, proving the effectiveness of the comprehensive hierarchical design.

(4) The goal of our OPEN is to measure critical local sub-spaces, while the intuitive approach is to apply weights to each dimension, i.e., directly using full connection (FC) layers. Thus, we experiment with the FC version of OPEN, which can be seen as inferior. It may be because when there is only one dimension to represent latent semantics, whose representation ability is minimal, which verifies the necessity and effectiveness of our sub-space-aware design.

(5) When the optimal combination sub-space-aware pattern mining is ignored, the performance degrades. Since not all size levels can observe effective latent characteristics, and the optimal state between each level cannot be mined, blindly using non-optimal levels may introduce interference.

(6) Based on the optimality prior of sub-space-aware patterns in Subsection 4.2.2, we only use the middle-level sub-spaces for training (i.e., $n = 4, 5, 6$). At this time, we can also find that the performance is improved compared to the baseline, indicating that this prior can also be used directly to improve performance without calculating all sub-space levels. Yet, this prior is only a simple validity verification, since it cannot guarantee the optimal combination performance of all potential size-levels.

(7) In contrast to the proposed discrete dynamic mining, we also investigate continuous weight learning that uses learnable weights to sum all size levels. As depicted in Figure 5(a) that shows all optional states of OPEN model, based on the dynamic searching results, our method can produce more concise and sparse results, while continuous weights have many non-zero values as Figure 5(b), which leads to subtle disturbance, e.g., 'OPEN (CW-learning)' is worse than 'OPEN (average)' in Table 4.

4.3.2 *Generality and efficiency*

(1) To further verify the generality of our optimal sub-space-aware learning, we apply the proposed OPEN to more representative models, including typical holistic matching methods VSE++ [2] and VSE∞ [3] (O2O-based), and the typical cross-attention methods SCAN [6], and BFAN [7] (M2M-based). As shown in Table 5, compared with the baselines, we observe that our OPEN can significantly improve the performance with respect to all evaluation

Table 5 Performance comparison of the OPEN framework on different baselines on Flickr30K and MS-COCO. *: Reproduced results using the same feature backbone for fairness. The results of proposed method are in bold.

Method		Flickr30K dataset					MS-COCO dataset				
		Image-to-text		Text-to-image		rSum	Image-to-text		Text-to-image		rSum
		R@1		R@10			R@1	R@10	R@1	R@10	
M2M-based	SCAN [6]	67.9	94.4	43.9	82.8	452.2	69.2	97.5	54.4	93.6	493.9
	+OPEN(g-)	**76.1**	**97.5**	**58.1**	**90.0**	**499.5**	**77.2**	**98.3**	**61.3**	**95.4**	**516.7**
	+OPEN(a-)	**78.6**	**97.8**	**58.3**	**89.9**	**504.0**	**77.7**	**98.6**	**61.9**	**95.5**	**519.2**
	BFAN [7]	64.5	–	48.8	–	280.3	70.9	–	58.8	–	312.9
	+OPEN(g-)	**77.6**	**98.0**	**58.8**	**90.2**	**503.7**	**77.9**	**98.9**	**61.7**	**95.6**	**518.6**
	+OPEN(a-)	**79.2**	**97.7**	**59.8**	**90.3**	**506.3**	**78.2**	**98.5**	**62.3**	**95.8**	**520.5**
O2O-based	VSE++ [2]	58.1	91.4	41.6	79.9	426.0	65.7	96.3	52.1	92.5	481.8
	+OPEN(g-)	**66.2**	**94.1**	**50.0**	**85.4**	**462.7**	**71.5**	**97.5**	**56.9**	**94.1**	**500.2**
	+OPEN(a-)	**66.9**	**93.1**	**49.7**	**85.1**	**460.1**	**72.3**	**97.8**	**57.3**	**94.7**	**503.6**
	VSE∞ [3]	76.5	97.7	56.4	89.9	498.1	78.5	98.7	61.7	95.6	520.8
	+OPEN(g-)	**78.4**	**97.9**	**58.8**	**90.9**	**505.4**	**80.2**	**98.7**	**64.1**	**96.2**	**526.6**
	+OPEN(a-)	**79.1**	**98.1**	**58.9**	**90.6**	**507.8**	**80.1**	**98.9**	**63.9**	**95.9**	**526.1**

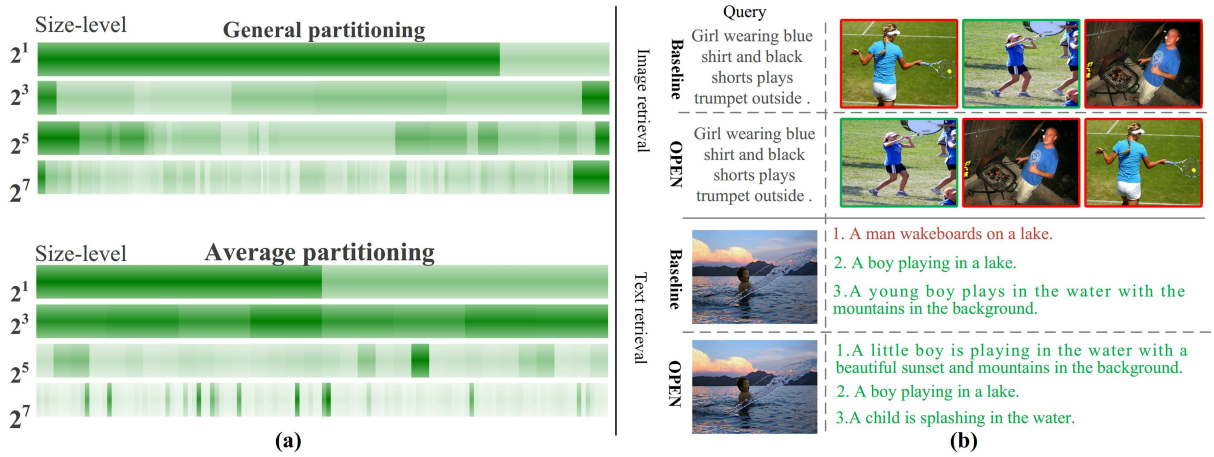

Figure 6 (Color online) (a) Visualization of the importance of different local sub-spaces (internal dimensions) at specific size-level patterns, e.g., 1, 3, 5, 7 on the OPEN model, where the darker the color, the greater the importance. (b) Retrieval cases comparison, which shows the top 3 ranked images and texts, where the correct ones are marked in green.

metrics, which demonstrates the superiority of our framework.

(2) The increased learning parameters in our OPEN framework are the MLP in (4), which are negligible compared with the baseline model. Specifically, with the fair settings (same batch size and running environments), the matching time required for the Flickr30K test set is at the same level as the baselines. For example, for OPEN based on VSE∞ [3], the average retrieval time is 0.9365×10^{-6} s, which is similar to that of [3], i.e., 0.2428×10^{-6} s. It verifies that our OPEN can achieve effective performance gains with a slight extra computation.

4.4 Visualization and case analysis

To better understand the effectiveness of our proposed model, we provide the visualization to show the learned weights for different size-level local sub-spaces, which are depicted in Figure 6(a). We show the cases corresponding to specific size-levels (e.g., 1, 3, 5, 7) in the best model of OPEN(g-) and OPEN(a-), respectively. Obviously, under the different size-level sub-space patterns, the local dimensions constitute different importance for similarity measurement, i.e., the dark green part within the whole feature. It verifies our key idea that measures the similarity between images and text from different size-level local sub-spaces, thus capturing the finer alignment of crucial latent characteristics within the features.

Specific image-text retrieval cases are shown in Figure 6(b). Compared with the baseline, in which cross-modal feature similarity is directly examined in the whole representation space, our OPEN can non-uniformly capture the important finer alignment between such local feature dimensions, thereby achieving more precise similarity retrieval between images and texts.

5 Conclusion

In this paper, we propose a novel OPEN. Different from existing methods, our OPEN devises the optimal learning pattern to further align finer local sub-spaces within the feature. Mainly, we propose optimal combined sub-space-aware patterns, which dynamically discover the optimal complementarities of different size-level sub-space-aware patterns. Besides, we also reveal that the pattern with the middle-size-level sub-spaces can achieve optimal performance gains with maximum probability. Extensive experiments demonstrate the superiority of simplicity, generality, and effectiveness of our OPEN framework. For future work, we intend to exploit this general framework for other vision-language tasks, such as video-text retrieval and visual grounding.

Acknowledgements This work was supported by Science Fund for Creative Research Groups (Grant No. 62121002), National Natural Science Foundation of China (Grant Nos. 62502490, 62222212, 62336001), Artificial Intelligence-National Science and Technology Major Project (Grant No. 2023ZD0121200), Natural Science Foundation of Jiangsu Province (Grant No. BK20250496), Jiangsu Funding Program for Excellent Postdoctoral Talent, and the China Postdoctoral Science Foundation (Grant No. 2024M763178).

References

1 Li Z, Guo C, Feng Z, et al. Multi-view visual semantic embedding. In: Proceedings of International Joint Conference on Artificial Intelligence, 2022. 7
2 Faghri F, Fleet D J, Kiros J R, et al. Vse++: improving visual-semantic embeddings with hard negatives. 2017. ArXiv:1707.05612
3 Chen J, Hu H, Wu H, et al. Learning the best pooling strategy for visual semantic embedding. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 15789–15798
4 Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. In: Proceedings of International Conference on Machine Learning, 2021. 8748–8763
5 Karpathy A, Joulin A, Fei-Fei L. Deep fragment embeddings for bidirectional image sentence mapping. In: Proceedings of International Conference on Neural Information Processing Systems, 2014. 1889–1897
6 Lee K H, Chen X, Hua G, et al. Stacked cross attention for image-text matching. In: Proceedings of European Conference on Computer Vision (ECCV), 2018. 201–216
7 Liu C, Mao Z, Liu A A, et al. Focus your attention: a bidirectional focal attention network for image-text matching. In: Proceedings of ACM International Conference on Multimedia, 2019. 3–11
8 Chen H, Ding G, Liu X, et al. IMRAM: iterative matching with recurrent attention memory for cross-modal image-text retrieval. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 12655–12663
9 Diao H, Zhang Y, Ma L, et al. Similarity reasoning and filtration for image-text matching. In: Proceedings of AAAI Conference on Artificial Intelligence, 2021. 1218–1226
10 Chen T, Luo J. Expressing objects just like words: recurrent visual embedding for image-text matching. In: Proceedings of AAAI Conference on Artificial Intelligence, 2020. 10583–10590
11 Qu L, Liu M, Wu J, et al. Dynamic modality interaction modeling for image-text retrieval. In: Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021. 1104–1113
12 Wei J, Xu X, Yang Y, et al. Universal weighting metric learning for cross-modal matching. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 13005–13014
13 Zhang K, Mao Z, Wang Q, et al. Negative-aware attention framework for image-text matching. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 15661–15670
14 Zhang K, Zhang L, Hu B, et al. Unlocking the power of cross-dimensional semantic dependency for image-text matching. In: Proceedings of ACM International Conference on Multimedia, 2023. 4828–4837
15 Gu J, Cai J, Joty S R, et al. Look, imagine and match: improving textual-visual cross-modal retrieval with generative models. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018. 7181–7189
16 Li K, Zhang Y, Li K, et al. Visual semantic reasoning for image-text matching. In: Proceedings of IEEE/CVF International Conference on Computer Vision, 2019. 4654–4662
17 Xu X, Lin K, Yang Y, et al. Joint feature synthesis and embedding: adversarial cross-modal retrieval revisited. IEEE Trans Pattern Anal Mach Intell, 2020, 44: 3030–3047
18 Wehrmann J, Kolling C, Barros R C. Adaptive cross-modal embeddings for image-text alignment. In: Proceedings of AAAI Conference on Artificial Intelligence, 2020. 12313–12320
19 Wang L, Li Y, Lazebnik S. Learning deep structure-preserving image-text embeddings. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016. 5005–5013
20 Zhou M, Niu Z, Wang L, et al. Ladder loss for coherent visual-semantic embedding. In: Proceedings of AAAI Conference on Artificial Intelligence, 2020. 13050–13057
21 Chen T, Deng J, Luo J. Adaptive offline quintuplet loss for image-text matching. In: Proceedings of European Conference on Computer Vision, 2020. 549–565
22 Zhang Y, Ji Z, Wang D, et al. USER: unified semantic enhancement with momentum contrast for image-text retrieval. IEEE Trans Image Process, 2024, 33: 595–609
23 Zhang K, Mao Z, Liu A A, et al. Unified adaptive relevance distinguishable attention network for image-text matching. IEEE Trans Multimedia, 2022, 25: 1320–1332
24 Liu C, Mao Z, Zhang T, et al. Graph structured network for image-text matching. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 10921–10930
25 Ji Z, Chen K, Wang H. Step-wise hierarchical alignment network for image-text matching. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, 2021, 765–771
26 Wu D, Li H, Gu C, et al. Dual stream relation learning network for image-text retrieval. IEEE Trans Multimedia, 2024, 27: 1551–1565
27 Zhang Y, Ji Z, Pang Y, et al. Hierarchical and complementary experts transformer with momentum invariance for image-text retrieval. Knowl-Based Syst, 2024, 309: 112912
28 Zhang K, Hu B, Zhang H, et al. Enhanced semantic similarity learning framework for image-text matching. IEEE Trans Circuits Syst Video Technol, 2024, 34: 2973–2988
29 Liu A A, Yang L, Li W, et al. Multi-level semantics probability embedding for image-text matching. Inf Process Manage, 2025, 62: 103968
30 Zhang H, Zhang L, Zhang K, et al. Identification of necessary semantic undertakers in the causal view for image-text matching. In: Proceedings of AAAI Conference on Artificial Intelligence, 2024. 7105–7114
31 Li Z, Zhang L, Zhang K, et al. Fast, accurate, and lightweight memory-enhanced embedding learning framework for image-text retrieval. IEEE Trans Circuits Syst Video Technol, 2024, 34: 6542–6558

32 Lu C, Zhang N, Sun S. A lightweight multi-grained image-text retrieval paradigm via cascaded representation learning and parameter-free feature aggregation. IEEE Trans Circuits Syst Video Technol, 2024, 34: 13584–13595
33 Huang Y, Wang L. Acmm: Aligned cross-modal memory for few-shot image and sentence matching. In: Proceedings of IEEE/CVF International Conference on Computer Vision, 2019. 5774–5783
34 Wang S, Wang R, Yao Z, et al. Cross-modal scene graph matching for relationship-aware image-text retrieval. In: Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision, 2020. 1508–1517
35 Wang S, Zhou F, Yang M, et al. SGG-MVAR: cross-modal retrieval with scene graph generation and multiview attribute relationship guidance. IEEE Trans Comput Soc Syst, 2025, 12: 3671–3683
36 Chen Y C, Li L, Yu L, et al. Uniter: universal image-text representation learning. In: Proceedings of European Conference on Computer Vision, 2020. 104–120
37 Gan Z, Chen Y C, Li L, et al. Large-scale adversarial training for vision-and-language representation learning. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2020. 6616–6628
38 Li W, Gao C, Niu G, et al. Unimo: towards unified-modal understanding and generation via cross-modal contrastive learning. In: Proceedings of Association for Computational Linguistics, 2021. 2592–2607
39 Jia C, Yang Y, Xia Y, et al. Scaling up visual and vision-language representation learning with noisy text supervision. In: Proceedings of International Conference on Machine Learning, 2021. 4904–4916
40 Li J, Selvaraju R, Gotmare A, et al. Align before fuse: vision and language representation learning with momentum distillation. In: Proceedings of Advances in Neural Information Processing Systems, 2021. 9694–9705
41 Zhu Y, Xu H, Du A, et al. Image-text matching model based on CLIP bimodal encoding. Appl Sci, 2024, 14: 10384
42 Hassan D, Domínguez J, Midtvedt B, et al. Cross-modality transformations in biological microscopy enabled by deep learning. Adv Photon, 2024, 6: 1–64
43 Guan Z, Zhao W, Liu H, et al. Cross-modal guided visual representation learning for social image retrieval. In: Proceedings of IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024. 1–14
44 Zhang H, Mao Z, Zhang K, et al. Show your faith: Cross-modal confidence-aware network for image-text matching. In: Proceedings of AAAI Conference on Artificial Intelligence, 2022. 3262–3270
45 Wang Y, Su Y, Li W, et al. Dual-path rare content enhancement network for image and text matching. IEEE Trans Circuits Syst Video Technol, 2023, 33: 6144–6158
46 Fu Z, Mao Z, Song Y, et al. Learning semantic relationship among instances for image-text matching. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 15159–15168
47 Pan Z, Wu F, Zhang B. Fine-grained image-text matching by cross-modal hard aligning network. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 19275–19284
48 Li K, Zhang Y, Li K, et al. Image-text embedding learning via visual and textual semantic reasoning. IEEE Trans Pattern Anal Mach Intell, 2023, 45: 641–656
49 Anderson P, He X, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018. 6077–6086
50 Krishna R, Zhu Y, Groth O, et al. Visual genome: connecting language and vision using crowdsourced dense image annotations. Int J Comput Vis, 2017, 123: 32–73
51 Pennington J, Socher R, Manning C. GloVe: global vectors for word representation. In: Proceedings of Conference on Empirical Methods in Natural Language Processing, 2014. 1532–1543
52 Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics, 2019. 4171–4186
53 Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of International Conference on Learning Representations, 2021
54 Song Y, Soleymani M. Polysemous visual-semantic embedding for cross-modal retrieval. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 1979–1988
55 Kim D, Kim N, Kwak S. Improving cross-modal retrieval with set of diverse embeddings. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 23422–23431
56 He S, Zhang Y, Xie R, et al. Rethinking image aesthetics assessment: models, datasets and benchmarks. In: Proceedings of International Joint Conference on Artificial Intelligence, 2022. 942–948
57 Karpathy A, Fei-Fei L. Deep visual-semantic alignments for generating image descriptions. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015. 3128–3137