# Momentum-Based Uni-Modal Soft-Label Alignment and Multi-Modal Latent Projection Networks for Optimizing Image-Text Retrieval

Xiaole Zhu[1] Zongtao Duan[2*]
Junchen Huang[3] Xing Sheng[4]
School of Information Engineering Chang'an University
Shanxi Province, China
{xiaolezhu[1],ztduan[2],jchuang[3]}@chd.edu.cn
xingS1992@126.com[4]

## Abstract

**Image-text retrieval (ITR) has made significant progress in recent years. However, it still faces two major challenges. The first challenge is the problem of intra-modal semantic loss issue, which is related to the lack of semantic associations between single-modal data. The second challenge is that different modal data cannot be effectively mapped to the same shared space, resulting in inconsistent representations between multimodal data, making it difficult to perform effective alignment and fusion. These challenges lead to limitations in the generality and retrieval accuracy of existing ITR retrieval models and challenges in the validity and reliability of these methods in practical applications. We propose two new methods to address these challenges: the Unimodal Momentum Soft Label Alignment (UMSA) method and the Multimodal Data Potential Projection (MMLP) method. Our methods aim to establish semantic links between unimodal data and overcome the underfitting problem of linearly mapping multimodal data into the same shared space. Our method has been extensively experimentally validated on various ITR models and datasets, all showing significant improvements in retrieval performance, including zero-sample retrieval performance. In addition, this approach is compatible with a wide range of ITR retrieval models, thereby improving model generality and accuracy.**

***Keywords:*** ***Cross-modal retrieval, Visual language model, Multimodal alignment, Zero-shot retrieval.***

## 1. Introduction

With the rapid development of intelligent information technology, multimodal data, including text, audio, video, and images, are ubiquitous in our daily lives. These diverse content forms not only enrich our life experience but also help us perceive and understand the world around us more comprehensively and accurately. Humans can easily align and complement different forms of information, allowing us to better learn and absorb knowledge. In the cross-modal field of artificial intelligence, the research goal is to achieve semantic alignment and complementary functions for different forms of information similar to the human brain. With the in-depth study of multi-modal technology, multimedia data can be interconnected. Through cross-modal alignment and fusion technology, one or more media data can complement each other and enhance their semantic information, thereby enabling computers to understand data information in multiple modalities more comprehensively.

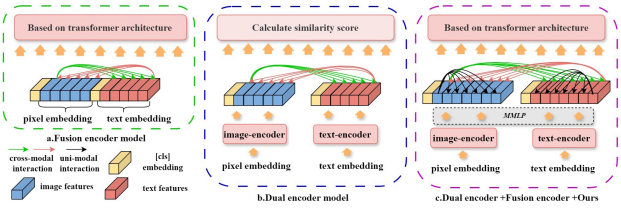



Figure 1: As shown in the figure, among the three existing classifications of visual language models, most models only consider cross-modal interaction and ignore single-modal interaction, while our model considers both.

Image-text retrieval (ITR) is a basic task among many multi-modal tasks. It uses computer vision and natural language processing technology to achieve bidirectional retrieval between images and text. The main goal is to extract information from massive images. Quickly and accurately retrieve related image and text information based on user queries in text data. It is mainly divided into image retrieval through text (IR), which means finding the image that is closest to the text description in the image pool, and the other is text retrieval through image (TR), which means finding the text that best describes the image in the text pool. In recent years, image text retrieval methods have achieved

*Corresponding Authors: Zongtao Duan (ztduan@chd.edu.cn)

near state-of-the-art (SOTA) performance [16, 15, 36]. Although these methods have achieved good retrieval results, the key challenge is to align data in different modalities and make up for semantic matching in different modal data channels. Good semantic correspondence directly affects the measurement of similarity between images and texts. However, current methods in the literature [1, 13, 16, 35] tend to focus on optimizing the alignment problem between multi-modal data. In contrast, the problem of aligning similar data within a single channel is often ignored. Although cross-modal data alignment is crucial in multimodal learning, small but significant differences can exist within a single channel, even for similar data. These differences may affect the model during data processing and feature extraction, thereby affecting its performance and generalization ability. Therefore, in addition to cross-modal data alignment, the alignment of similar data within a single channel also deserves in-depth study. By solving this problem, we can more effectively mine relevant information within the data, thereby improving the model's generalization.

In addition, the features extracted by image and text encoders are usually embedded in separate semantic spaces. This difference originates from factors such as different semantic structures between modalities and the perceptual characteristics of the data. Because feature representations between different modalities are often semantically different and heterogeneous, direct interaction and modeling may increase the similarity between positive samples and reduce the model's generalization ability. Therefore, effectively establishing cross-modal interaction and fusion mechanisms in different semantic spaces has become essential in designing multi-modal encoders. As shown in a) and b) in Figure. 1, in previous works, whether a single-stream structure or a dual-stream structure, they focused on the interaction between cross-modalities and used the similarity score as the criterion for alignment. However, the alignment problem between single modalities must be addressed, leading to the model's poor generalization ability and, thus, our work's primary motivation. This paper proposes a new dual-stream structure method (UMSA), as shown in c) in Figure. 1, based on single-modal alignment, which uses momentum soft labels to guide the ITR model to perform single-modal data alignment. This method achieves data alignment between cross-modalities, effectively identifies similar samples in single modalities, and distinguishes similar samples so that positive sample pairs with higher similarity scores can be compared with images and images. Samples with higher similarity between texts are aligned in the same shared space. In addition, among the methods of mapping modalities into the same space, we propose the cross-modal data latent mapping method (MMLP) to represent different modal data in the same shared space. It can learn the nonlinear interaction and combination of different input features through multiple hidden layers, generate high-order feature representations, have high data adaptability, and better generalize to unseen data. We conducted many experiments on different ITR models and data sets, proving that our method can effectively improve image and text retrieval performance. Our method can improve performance by 1.2% to 3.9% on the RSUM metric on different benchmark models. Our main contributions are summarized as follows:

- We designed a retrieval method UMSA that takes into account single-modal alignment, using the momentum soft labels generated by the teacher model to guide the alignment of single-modal data;
- We propose the MMLP method to map different modal data into the same space, which improves the general generalization ability of the model and the interaction performance between modalities;
- We have conducted a large number of experiments on different ITR models and data sets, and the results show that our method can effectively improve the performance of image and text retrieval models and achieve overall performance better than the baseline model.

## 2. Related Works

### 2.1. Image-Text Retrieval

Image-text retrieval combines image and text data to provide comprehensive and accurate information retrieval. It enhances the quality of retrieval results by providing richer semantic information and visual features. The challenge is to achieve cross-modal semantic understanding and similarity matching. Current methods rely on target detectors [14, 31, 32] to extract entities and their regions in the image. Then, they use multi-modal encoders to align and fuse text and image features. However, these methods may lead to information loss, the inability to handle modal imbalance, ignoring correlation and semantic information between different modalities, and difficulty processing heterogeneous data. Due to different characteristics, data distribution, and feature spaces, some modal features may be over-emphasized or ignored, making it challenging to capture the correlation and semantic information between different modalities, thus reducing the accuracy and effectiveness of retrieval. Some methods use a dual-stream encoder structure, where an image encoder and a text encoder extract features of images and texts, respectively. Then, they use a method similar to contrastive learning [34] for alignment and fusion. This method efficiently calculates similarity scores, making retrieval speed and efficiency efficient. However, more effective information interaction and

fusion between modalities during the entire retrieval process is needed, which may result in the information complementarity between modalities not being fully utilized, and it is challenging to handle the heterogeneity between modalities. Other methods use a dual-stream encoder + a multimodal fusion encoder to extract text and image features from different encoders. They then use the transformer's self-attention mechanism fusion encoder to perform modal alignment and fusion. These models use complex labels as supervision signals to guide the model's training process. However, they only align the annotated images and texts in the dataset, ignoring the potential semantic correlation between different images and texts.

Our method uses the external soft labels provided by the teacher model of momentum self-distillation (that is, the student model gradually evolves into the teacher model as the training progresses, and the parameter ratio of the teacher to the student model is 0.995), that is, the model learns from the data through the algorithm, and as a result, it can represent richer semantics. It can capture the implicit information between labels. It can make the feature representation between different modalities more transparent, interpretable, and numerically consistent, reducing the differences between modalities. Imbalance, which is conducive to cross-modal alignment to improve the model's generalization ability, understand the model's working mechanism and decision-making process in different modalities, and achieve better retrieval performance.

### 2.2. Multimodal Data Mapping and Alignment

Multi-modal learning transforms data from different sources into a unified representation, facilitating more efficient processing and analysis. This shared representation helps uncover correlations and common features between various modalities, making cross-modal information comparable and analyzable within a unified representation space. The primary goal is to map data from different modalities into a consistent representation, simplifying subsequent tasks and reducing complexity. However, most existing Image-Text Retrieval (ITR) models [23, 11] rely on linear mapping techniques, which are limited in their ability to capture the rich, complex information embedded in image and text data. These data often contain multi-scale, multi-directional features and semantic and contextual information, which linear mappings fail to exploit fully.

To address this limitation, we propose MMLP as a key enhancement to the network architecture. MMLP leverages multiple nonlinear layers and activation functions, enabling the model to perform nonlinear transformations that can better capture the inherent complexity of the data. This allows for more flexible and adaptive feature extraction, adjusting model complexity based on the data's characteristics.In contrast to linear mapping, which can only model simple, linear relationships, MMLP captures both linear and nonlinear dependencies, resulting in significantly improved model expressiveness and generalization ability. Our experiments consistently show that MMLP outperforms traditional linear methods by effectively modeling the intricate patterns within multi-modal data, leading to superior retrieval performance in cross-modal tasks.

## 3. Proposed Method

### 3.1. Image-text Contrastive Learning

Image-text contrastive learning is a method for multimodal learning that aims to exploit the correlation between images and text to improve model performance. This method uses a contrastive loss function to learn the semantic correlation between images and texts by bringing the same sample's different modal representations closer and pushing different samples apart. Specifically, image-text comparison learning maps image and text representations into a shared feature space so that similar images and texts are closer in the feature space while dissimilar photos and texts are further apart. The data in the data set exists in the form of image-text pairs $\{(I_i, T_i)\}_{i=1}^{N}$, where $(I_i, T_i)$ represents the relationship between image and text as a sample pair. Contrastive learning-based methods [9, 34] align these image-text pairs. Specifically, multiple image-text pairs are sampled from the dataset to form a batch according to batch size. During training, the distance between correctly matched images and text pairs in the feature space is continuously drawn closer. In contrast, the distance between unpaired images and text pairs in the feature space is constantly drawn away. We maintain two queues to store the most recent M image-text representations from the momentum unimodal encoders. First, we define similarity in the following way:

$$s\left(I,T\right)=g_v\left(v_{cls}\right)^T g_w\left(w_{cls}\right) \tag{1}$$

where $g_v$ and $g_w$ are linear transformations that map the [CLS] embeddings to normalized lower-dimensional (256-d) representations, $v_{cls}$ and $w_{cls}$ are the output [CLS] embeddings of the visual encoder and text encoder respectively. And are transformations that map [CLS] embeddings to normalized low-dimensional representations. Based on this, we calculate the similarity between softmax normalized images and texts within batches with the following formula:

$$p_m^{i2t}\left(I\right)=\frac{exp\left(s\left(I,T_m\right)/\tau\right)}{\sum_{m=1}^{M}exp\left(s\left(I,T_m\right)/\tau\right)} \tag{2}$$

$$p_m^{t2i}\left(T\right)=\frac{exp\left(s\left(T,I_m\right)/\tau\right)}{\sum_{m=1}^{M}exp\left(s\left(T,I_m\right)/\tau\right)} \tag{3}$$

where $\tau$ is a learnable temperature parameter. Assume that $y^{i2t}(I)$ and $y^{t2i}(T)$ represent the one-hot similarity of the real label, in which only the probability of positive sample pairs is 1, and the probability of other samples is 0. Finally, the contrastive loss is defined as p and y, and the cross-entropy loss between is H:

$$L_{itc} = \frac{1}{2}\mathbb{E}_{(I,T)\sim D}\left[H(y^{i2t}(I), p^{t2i}(I)) + H(y^{t2i}(T), p^{t2i}(T))\right] \quad (4)$$

where $\mathbb{E}_{(I,T)\sim D}$ represents the average loss of the model on a given data distribution D, that is, the average performance of the model on different image-text pairs.

### 3.2. Image-Text Matching

Image-text matching is a multimodal learning method that aims to achieve semantic matching between images and text and predict whether the image and text are a matching or unpaired pair. We use the [CLS] token embedding of the multimodal encoder as the joint representation of image and text input it into the $itm_{head}$ binary classifier, and finally predict the two categories by splicing a fully connected (FC) layer and softmax activation function. The probability of its ITM loss is expressed as:

$$L_{itm} = \mathbb{E}_{(I,T)\sim D} H(y^{itm}, p^{itm}(I,T)) \quad (5)$$

where $y^{itm}$ is the two-dimensional one-hot representation representing the true label. We use the loss function of the original ITR model, which is

$$L_{original} = L_{itc} + L_{itm} \quad (6)$$

### 3.3. Single-modal Momentum Soft Label Alignment

Although good results have been achieved in image and text retrieval tasks, unimodal alignment is ignored in these models, which may affect the model's generalization performance to unknown data. As shown in the Figure. 2, we use t-distributed stochastic neighbor embedding (t-SNE)[29] to map images and text in high-dimensional space to three-dimensional space while maximizing the At this time, the similarity between the image and the text can be intuitively displayed as the distance relationship between points in the three-dimensional space in the t-SNE diagram. Among them, Image 1, Text 1, Image 3, and Text 3 are samples in the training set, and Image 2 and Text 2 are data that we have yet to see during training. Most existing ITR retrieval models can well align the three image-text pairs in Figure. 2. However, (a) due to the lack of information interaction between image and image, text and text, the two image-text pairs are mapped in the same space in different areas. When we use the same image and text encoder to extract features from image and text pairs, image 2 is closer to 1 at the pixel feature level, so it is mapped to an area adjacent to Text 2, which is closer to Text 3 regarding text features, so Text 2 is mapped to the adjacent area of Text 3. Therefore, a single-modal momentum soft label alignment method is introduced in our work, aiming to utilize the complementary information between images and text to improve the semantic correlation between them through momentum update. The method first utilizes a single-modal image encoder and a text encoder to convert images and texts into feature representations. Then, iteratively updates these feature representations through momentum updates to capture the semantic associations between them. In each update process, known image-text pairs are converted into label vectors by generating soft labels containing the probability that the image or text belongs to each category. Then, the generated soft labels are used as supervision signals to maximize the similarity between the image and text by aligning their representations, and the model parameters are updated through the backpropagation algorithm to improve the model's performance. As shown in Figure. 3, we first obtain the image feature $I'$ and text feature $T'$ from the teacher model of momentum distillation, then get $i'$ and $t'$ through the MMLP mapping and calculate the cosine similarity between $I'_i$ and $I'_j$ as $s^{i2i}_{ij}$, S and the similarity between $T'_i$ and $T'_j$ as $s^{t2t}_{ij}$. After that, the softmax normalized image-to-image $Q^{i2i}_{ij}$ and text-to-text $Q^{t2t}_{ij}$ similarities are obtained:

$$Q^{i2i}_{ij} = \frac{exp(S^{i2i}_{ij})/\tau}{\sum_{j=1}^{N} exp(S^{i2i}_{ij})/\tau} \quad (7)$$

$$Q^{t2t}_{ij} = \frac{exp(S^{t2t}_{ij})/\tau}{\sum_{j=1}^{N} exp(S^{t2t}_{ij})/\tau} \quad (8)$$

Secondly, we obtained the student models, namely ITR retrieval models $P^{i2i}_i$ and $P^{t2t}_i$ in the above way. We denote the probability distribution($Q^{i2i}_{i1}$,...,$Q^{i2i}_{iN}$) as $Q^{i2i}_i$ and use the similar step to obtain $Q^{t2t}_i$. During the training process, we used KL divergence to use $Q^{i2i}_i$ and $Q^{t2t}_i$ respectively to guide the alignment loss of single-modal data. This alignment loss can effectively promote Alignment between uni-modal data. Taking $Q^{i2i}_i$ as the target distribution, KL divergence is used to guide the learnable distribution $P^{i2i}_i$ for image-text alignment. Meanwhile, taking $Q^{t2t}_i$ as the target distribution, considering that the distributions

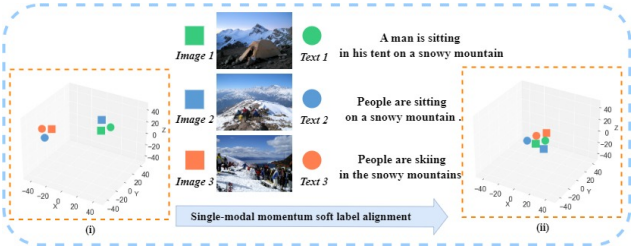



Figure 2: Ignore the data aligned within the modality and the feature distribution in the reduced dimension space after adding UMSA;

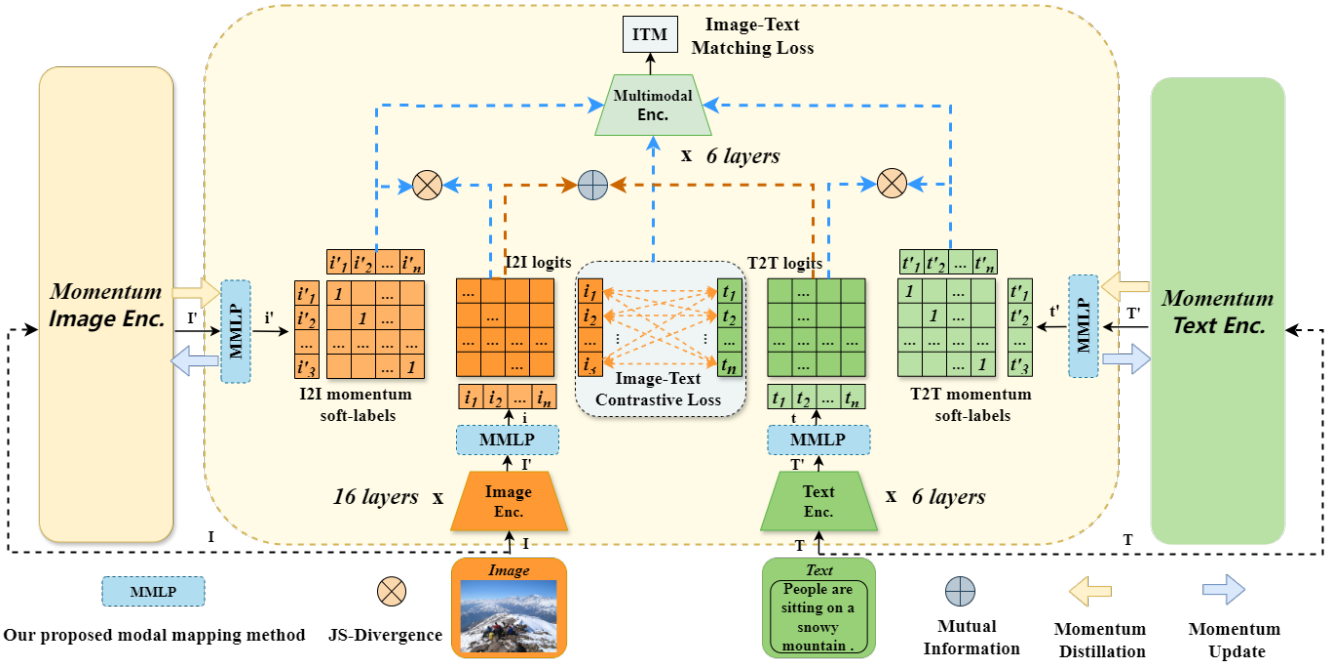



Figure 3: Illustration of our approach. It consists of an image encoder, text encoder, and multi-modal encoder, as well as a momentum teacher model. We propose a single-channel data alignment loss that uses soft labels generated by the momentum teacher model as additional supervision during training to adjust the unimodal representation of image-text pairs before fusion. To improve the shortcomings of linear mapping of multi-modal data, we use the MMLP data latent mapping method to make the model have better data adaptability during the training process.

P and Q may have the case that the probability of the sample points is zero, which will lead to the instability of the calculation of the KL divergence, and make his value become infinite. Therefore, we adopt a symmetric form of KL divergence, Jensen-Shannon divergence (JSD) [7], to bootstrap the learnable distribution $P_i^{t2t}$ for text-image alignment, which ensures that the computed values are more stable and finite. For the image modality, we define its JSD divergence as:

$$JSD(Q_i^{i2i} \parallel P_i^{i2i}) = \frac{1}{2}\left(D_{KL}(Q_i^{i2i}||M_i^{i2i}) + D_{KL}(P_i^{i2i}|M_i^{i2i})\right) \quad (9)$$

Among them, $M_i^{i2i} = \frac{1}{2}(Q_i^{i2i} + P_i^{i2i})$ is a mixed distribution of $Q_i^{i2i}$ and $P_i^{i2i}$.

Similarly, for textual modalities, the JSD divergence is set to be:

$$JSD(Q_i^{t2t} \parallel P_i^{t2t}) = \frac{1}{2}\left(D_{KL}(Q_i^{t2t}||M_i^{t2t}) + D_{KL}(P_i^{t2t}|M_i^{t2t})\right) \quad (10)$$

Among them, $M_i^{t2t} = \frac{1}{2}(Q_i^{t2t} + P_i^{t2t})$ is a mixed distribution of $Q_i^{t2t}$ and $P_i^{t2t}$. The final loss function of UMSA is represented as follows:

$$L_{UMSA} = (JSD(Q_i^{i2i}||P_i^{i2i}) + JSD(Q_i^{t2t}||P_i^{t2t}))/2 \quad (11)$$

### 3.4. Multimodal Data Mapping and Alignment

Multimodal data representation is the process of integrating data from various sources into a unified data representation space. This shared space enables data from different modalities to be analyzed and processed uniformly, leading to cross-modal information fusion, alignment, and analysis. However, the model can only learn linear combinations of input features for linear mapping, depicted in Figure. 4. It tends to be inflexible and adaptable to data, resulting in underfitting problems, especially when dealing with complex data. As a result, the model's generalization ability could be much improved, ultimately leading to poor performance.

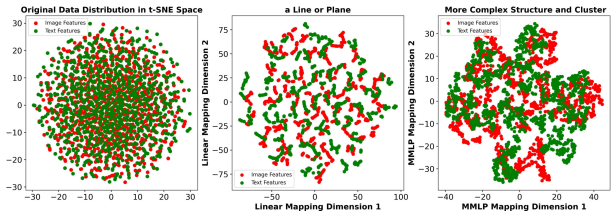



Figure 4: After the image and text data undergo linear and MMLP mapping, respectively, the t-SNE dimensionality reduction method is used to draw a scatter plot.

$$Z_X = WX + b_X, Z_Y = WY + b_Y. \quad (12)$$

Among them, $W$ is a linear weight matrix $b_X$ and $b_Y$ are bias vectors. $Z_X$ and $Z_Y$ are the linear representations of $X$ and $Y$ in $Z$ space respectively. Therefore, the shape of the image and text data distributed in the joint space after linear mapping is similar to a straight line or a plane in Figure. 4. The method we proposed, MMLP, can dynamically adjust the network structure and parameters according to the complexity and pattern of the data and can learn non-linear relationships and feature representations in image and text data, thereby making the distribution of data points in the shared space more accurate. And rich, able to better capture the complex structures and relationships between image data and the semantic structures and correlations between text data. After using the MMLP method to project different modal data into the same public space, the data formula is expressed as follows and distributed as follows:

$$Z_X = ReLU(f(W_3 \cdot ReLU(f(W_2 \cdot ReLU(f(W_1^X \cdot X)) + b_{X2})) + b_{X1})) \tag{13}$$

$$Z_Y = ReLU(f(W_3 \cdot ReLU(f(W_2 \cdot ReLU(f(W_1^Y \cdot Y)) + b_{Y2})) + b_{Y1})) \tag{14}$$

Among them, $f$ is the activation function, $W_1^X$ and $W_1^Y$ is the weight matrix of each modality $W_3$, $W_2$are the weight matrices of the hidden layer and output layer of the MMLP method respectively, $b_{X1}$ and $b_{X2}$ are the bias of text data mapping $b_{Y1}$ and $b_{Y2}$ are the bias of image data mapping.$Z_X$ and $Z_Y$ are the nonlinear mappings[26] of $X$ and $Y$ in $Z$-space, and are also the feature representations of image and text data. As shown in Figure. 4, we use part of the experimental data to use the MMLP nonlinear mapping method to map the data in the shared space with a similar shape to the clustering, which shows that the method can better preserve the high-dimensional nonlinear features of images and text.

Our approach, MMLP, introduces nonlinear properties using activation functions that can compress the input feature vectors layer by layer, enabling the model to capture complex intra- and inter-modal relationships. Nonlinear mapping can handle complex intra-modal feature transformations layer by layer, ultimately mapping features from different modalities into a high-dimensional shared space. This mapping aims to align image and text feature representations as much as possible so that their similarity in a unified space can reflect semantic correlations. Secondly, images and text often contain different semantic information in multimodal tasks. Images primarily convey visual content, while text provides semantic details and contextual information. The information of these two modalities is complementary and can provide richer content for the retrieval task through a reasonable alignment strategy. The nonlinear mapping of MMLP helps to integrate this complementary information effectively, making the representation of image and text in the public space more compatible, thus enhancing retrieval performance.

In this paper, we analyze the key role of the MMLP method in improving the performance of multimodal retrieval using information entropy and mutual information from the theory of information complementarity. We can define a generic formula for information entropy in Eq. 15 and mutual information [22] in Eq. 16 as:

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) \tag{15}$$

$$\mathrm{Info}(h_{\mathrm{img}}, h_{\mathrm{text}}) = I(h_{\mathrm{img}}; h_{\mathrm{text}}), \tag{16}$$

$$I(h_{\mathrm{img}}; h_{\mathrm{text}}) = H(h_{\mathrm{img}}) + H(h_{\mathrm{text}}) - H(h_{\mathrm{img}}, h_{\mathrm{text}}). \tag{17}$$

$$H(X,Y) = H(x) + H(Y|X) = -\sum_{x \in X} \sum_{y \in Y} P(x,y) \log P(x,y) \tag{18}$$

Among them, $I(h_{img}; h_{text})$ in Eq. 17 represents the amount of mutual dependence or shared information between image modality and text modality, which measures the degree of dependence between image modality and text modality. The larger the mutual information is, the stronger the relevance of image and text modalities in the public space, and the more fully complementary the information is. $H(h_{img})$ and $H(h_{text})$ in Eq. 17 and Eq. 16 represents the entropy of each of the image modality and text modality, which indicates the amount of information they each contain, and measures the dispersion or uncertainty of the representations within the modality; higher entropy values imply that the feature representations are too dispersed to extract the key information efficiently. $H(h_{img}, h_{text})$ represents joint entropy, which represents the amount of joint information of image and text modalities, measures the common uncertainty between the two modalities, and the joint entropy between the two modalities can be reduced by decreasing the entropies $H(X)$ and $H(Y)$ within the modalities. $H(Y|X)$ in Eq. 18 is the conditional entropy, which represents the uncertainty of the modality $Y$ given the modality $X$. This measures the amount of unexplained information remaining for modality $Y$ after the information about $X$ is known.

Firstly, MMLP compresses the feature representation layer by layer through the nonlinear activation function (ReLU) to remove unnecessary, redundant information, which makes the features within the modality more concentrated and the entropy value within the modality reduced. The entropy values of $H(X)$ and $H(Y)$ are reduced. The lower entropy value means that the model better extracts the key information in the modality, which is more effective in cross-modal alignment. Compared to linear mapping, MMLP captures linear relationships and extracts and expresses key nonlinear features in different modes more efficiently, significantly improving model performance in cross-modal tasks.

Secondly, the information complementarity between image and text modalities is enhanced by the mapping of the MMLP, which means that the MMLP can better remove redundant features and reduce irrelevant information, resulting in a reduction of the joint uncertainty (i.e., joint entropy) between $X$ and $Y$. The MMLP can also remove the redundant features and reduce the irrelevant information. By mapping the image modality $h_{img}$ and the textual modality $h_{text}$ into the same shared space, the MMLP can capture correlated features between the modalities. This mapping can be achieved by maximizing mutual information, and MMLP mapping makes the common information (dependencies) between modalities much stronger, which leads to a faster reduction of joint uncertainty. That is, the mutual information in Eq. 17 is enhanced by reducing the joint entropy between modalities $H(h_{img}, h_{text})$. At the same time, the representations of the image and text modalities are compressed and simplified through nonlinear mapping, preserving the most relevant features.

The role of MMLP is to reduce the joint entropy $H(X, Y)$ by simultaneously decreasing the intra-modal and conditional entropy, maximizing the mutual information between the image and text modalities $I(h_{img}; h_{text})$; and ensuring that the inter-modal semantic information is adequately complemented and fused. This complementarity enables the model to align and retrieve relevant content more accurately based on the shared representation space when performing image and text retrieval. We maximize the mutual information between image and text modalities as part of the loss of our model with the following equation:

$$\mathcal{L}_{\mathrm{MI}} = -I(h_{\mathrm{img}}; h_{\mathrm{text}}) = H(h_{\mathrm{img}}, h_{\mathrm{text}}) - H(h_{\mathrm{img}}) - H(h_{\mathrm{text}}) \tag{19}$$

### 3.5. Training Objective

We adjust the original loss of the ITR model using the UMSA loss, so the total loss function is expressed as:

$$L_{loss} = L_{original} + \alpha \cdot L_{UMSA} + \beta \cdot L_{MI} \tag{20}$$

Among them, $\alpha$ is the proportional coefficient. At first, we used meta-learning [8] to determine the coefficient $\alpha$ and $\beta$, but since meta-learning is suitable for scenarios with high task diversity and high requirements for fast adaptation to new tasks, and the multi-task division in our graphic retrieval is relatively single, we divide it into pairs of graphic with different categories and difficulties, which leads to the meta-learning algorithm not being able to take full advantage of its strengths. Eventually, we also proved in the experimental results that the parameters obtained by meta-learning are unsuitable for the task, so we used grid search instead of meta-learning to adjust the parameter $\alpha$. In our many experiments, it has been proved that when $\alpha$ is 0.6, and $\beta$ is 0.4, the model can achieve the best performance index. The approximate running flow of the whole model, as in Alg. 1.

**Algorithm 1** Momentum-based unimodal alignment and MMLP mapping of one-round cycles

**Data:** Student model's image and text features, momentum coefficient $\alpha = 0.995$, queue size $Q\_size = 57600$

**1. Map to common space via MMLP:**
$Q\_img \leftarrow mmlp(Q\_img)$ $Q\_txt \leftarrow mmlp(Q\_txt)$
$img\_feat \leftarrow mmlp(img\_feat)$
$txt\_feat \leftarrow mmlp(txt\_feat)$

**2. Momentum update for image and text features:**
$h\_img \leftarrow \alpha \cdot Q\_img + (1-\alpha) \cdot img\_feat$
$h\_txt \leftarrow \alpha \cdot Q\_txt + (1-\alpha) \cdot txt\_feat$

**3. Compute intra-modality (i2i, t2t) similarities:**
$sim\_i2i \leftarrow cos\_sim(h^i\_img, h^j\_img)$
$sim\_t2t \leftarrow cos\_sim(h^i\_txt, h^j\_txt)$

**4. Jensen-Shannon divergence for unimodal alignment:**
$jsd\_i2i \leftarrow JSD(softmax(sim^i\_i2i), softmax(sim^j\_i2i))$
$jsd\_t2t \leftarrow JSD(softmax(sim^i\_t2t), softmax(sim^j\_t2t))$

**5. Mutual information loss:**
$mi\_loss \leftarrow I(h\_img, h\_txt)$

**6. Final loss:**
$final\_loss \leftarrow loss_{original} + \alpha * (jsd\_i2t + jsd\_t2i) + \beta * mi\_loss$

**7. Update model parameters:**
$opt.step(final\_loss)$

**8. Momentum queue update:**
$Q\_img \leftarrow \alpha \cdot Q\_img + (1-\alpha) \cdot img\_feat$
$Q\_txt \leftarrow \alpha \cdot Q\_txt + (1-\alpha) \cdot txt\_feat$

**9. Ensure constant queue size:**
**if** *len(*$Q\_img$*)* $> Q\_size$ **then**
  Remove oldest image feature $Q\_img.pop(0)$
**if** *len(*$Q\_txt$*)* $> Q\_size$ **then**
  Remove oldest text feature $Q\_txt.pop(0)$

## 4. Experiments

### 4.1. Experimental Setup

#### 4.1.1 Datasets

We use the public datasets Flicker30k[33] and MSCOCO[19]. Flickr30k is an image uploaded by users of the Flickr image-sharing platform. Each image has 5 manual annotations. Image description covers the objects, scenes, and relationships between them that appear in the image. It contains a total of 30 thousand images and provides a total of 150 thousand images description annotations. MSCOCO is a large-scale image dataset released by Microsoft. The images in this dataset cover various scenes, and the text descriptions related to the images are more detailed and diverse. Each image contains at least 5 different text descriptions. Containing 10 thousand images,

approximately 50 thousand image text descriptions are provided. These datasets provide rich training data for image understanding and language-related tasks.

#### 4.1.2 Baselines

To evaluate the performance of our proposed method, we compare it with other ITR retrieval models. Most existing ITR retrieval models prioritize the information complementarity between cross-modalities and ignore data alignment within single-modal channels, achieving the best performance currently. We follow the baseline model's setup and optimization methods to ensure fair performance comparisons between models. Below, we introduce three typical models of existing ITR retrieval model architectures mentioned in related work, and these models are also used in our comparative experiments.

**(1) Fusion encoder ITR model** UNITER[4] and OSCAR[18] are single-stream pre-trained models for cross-modal representation learning; image and text data are processed simultaneously in one model and learned to encode them into a shared embedding space in the pre-training stage. This structure enables better capture of the semantic consistency between images and text, achieving good performance in various cross-modal tasks.

**(2) Dual-encoder ITR model** CLIP[26], FLIP[12], and LTBN [31] are three common two-stream models that use natural language supervision to facilitate transferable visual model learning. These models pre-train visual models using large-scale text data, allowing them to acquire rich visual knowledge. The models employ a separate image encoder and text encoder. They directly calculate the similarity between the image and text outputs to determine whether they match based on a given similarity threshold. This approach results in faster model response times but can lead to simple image and text alignment modeling. Among them, CLIP is a control experiment for our zero-shot retrieval experiment.

**(3) Dual encoder + Fusion encoder ITR model** ALBEF[17] and BLIP[15] are innovative pre-training methods that have been designed to improve visual-language understanding and generation. These methods use a unique approach that involves aligning visual and language models trained separately into a shared semantic space using the momentum distillation mechanism. The models are trained iteratively, with various loss functions, such as adversarial and similarity loss, used to guide the learning process. These methods aim to help the models achieve semantic consistency and alignment in cross-modal tasks, enabling them to learn more consistent and stable representations. This process can significantly improve the performance of the models on various tasks related to image and text understanding. During training, the models are guided by an adversarial loss function that helps them generate more realistic images. In contrast, the similarity loss function encourages the models to learn more meaningful semantic representations. Additionally, the momentum distillation mechanism is used to align the visual and language models to allow both models to improve their accuracy and consistency over time.

### 4.2. Implementation Details

For all retrieval experiments, we use the AdamW [21] optimizer with a base learning rate of 5e-6, weight decay of 5e-2, and cosine decay [20] to zero for the rest of training. Considering the trade-off between performance and model size and the model's similarity calculation score for images and texts depends on whether the features we extract are rich and diverse. We use an improved Vision Transformer(vit) [6] network on the image encoder side. Although the vit network performs well on image classification and other visual tasks, since the self-attention mechanism is not explicitly designed to capture local information but processes the entire image through a global attention mechanism, it has problems processing local information. There is a lack of explicit modeling of spatial structure and difficulty in capturing the spatial relationships between pixels. Inspired by the article[28], we slightly modified the patch layer of the vit[6] network. We used convolution channels and pooling operations to replace the native patch layer so that it can process local information while processing global information and retaining the image. Spatial information helps the model better understand the spatial relationships between pixels in the image. After that, we pretrained the modified vit network on Imagenet-1k[27]. We performed 25 thousand pre-training steps on the modified vit network on 8 A800 machines. The batch size was 128, and the final top1 accuracy Reached 86.4%. For the text encoder, we use the first six layers of the bert-base-uncased[5] tokenizer network fine-tuned from the pre-trained BERT to encode the text input. Regarding the multi-modal fusion encoder, we use the last six layers of bert-base-uncased as an encoder for the correlation and information complementation of image and text features. It is mainly used for the binary classification task of image and text matching (ITM).

### 4.3. Main Results

As shown in Table. 1, we compared different model methods for information-theoretic retrieval (ITR) models and evaluated their effectiveness in terms of R@K, a popular valuation index. We experimented with three architecture types of ITR models and found that our proposed method outperformed all the baseline models. Our proposed method effectively addresses the challenges of se-

Table 1: Comparison with state-of-the-art image-text retrieval methods, experimental results of image-text retrieval on MSCOCO and Flickr30K. Bold indicates best overall performance, for all other models, the paper's best results are reported regardless of model size/variables.

| | MSCOCO(5K test set) | | | | | | | Flickr30k(1K test set) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | TR | | | IR | | | | TR | | | IR | | | |
| Method | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | RSUM | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | RSUM |
| ***Single-stream ITR retrieval model*** | | | | | | | | | | | | | | |
| UNITER | 65.7 | 88.6 | 93.8 | 52.9 | 79.9 | 88.0 | 410.9 | 87.3 | 98.0 | 99.2 | 75.6 | 94.1 | 96.8 | 551.0 |
| VILT-B/32[13] | 61.5 | 86.3 | 92.7 | 42.7 | 72.9 | 83.1 | 439.2 | 83.5 | 96.7 | 98.6 | 64.4 | 88.7 | 93.8 | 525.7 |
| OSCAR | 70.0 | 91.1 | 95.5 | 54.0 | 80.8 | 88.5 | 479.9 | - | - | - | - | - | - | - |
| ALIGN[12] | 77.0 | 93.5 | 96.9 | 59.9 | 83.3 | 89.8 | 500.4 | 95.3 | 99.8 | 100.0 | 84.9 | 97.4 | 98.6 | 576.0 |
| **+Our method** | **77.8** | **94.0** | **97.3** | **61.1** | **83.9** | **90.1** | **504.2** | **95.9** | **99.9** | **100.0** | **85.4** | **97.9** | **99.0** | **578.1** |
| ***Dual-encoder ITR model*** | | | | | | | | | | | | | | |
| SCAN [14] | 50.4 | 82.2 | 90.0 | 38.6 | 69.3 | 80.4 | 410.9 | 67.4 | 90.3 | 95.8 | 48.6 | 77.7 | 85.2 | 465 |
| VSRN | 53.0 | 81.1 | 89.4 | 40.5 | 70.6 | 81.8 | 416.4 | 71.3 | 90.6 | 96.0 | 54.7 | 81.8 | 88.2 | 482.6 |
| FLIP | 60.2 | 82.6 | 89.9 | 44.2 | 69.2 | 78.4 | 424.5 | 89.1 | 98.5 | 99.6 | 75.4 | 92.5 | 95.9 | 551 |
| CPRD [3] | 70.8 | 91.7 | 96.2 | 53.4 | 80.6 | 88.5 | 481.2 | 90.7 | 99.0 | 99.7 | 78.6 | 94.9 | 97.4 | 560.3 |
| $VLMO_{base}$ [1] | 74.8 | 93.1 | 96.9 | 57.2 | 82.6 | 89.8 | 494.4 | 92.3 | 99.4 | 99.9 | 79.3 | 95.7 | 97.8 | 564.4 |
| **+Our method** | **75.2** | **93.5** | **97.3** | **57.8** | **82.9** | 89.8 | **496.5** | **92.9** | **99.6** | 99.9 | **80.5** | **96.1** | **98.2** | **567.2** |
| ***Dual encoder + Fusion encoder ITR model*** | | | | | | | | | | | | | | |
| MVSEN [25] | 58.7 | 84.0 | 91.7 | 42.5 | 72.0 | 82.7 | 431.6 | 81.7 | 95.6 | 98.2 | 63.1 | 88.0 | 92.9 | 519.5 |
| HGFN [24] | 76.4 | 95.2 | 98.2 | 62.8 | 90.3 | 95.6 | 518.5 | 75.3 | 94.2 | 97.2 | 57.4 | 83.1 | 89.6 | 496.8 |
| ALBEF | 77.6 | 94.3 | 97.2 | 60.7 | 84.3 | 90.5 | 504.6 | 95.9 | 99.8 | 100.0 | 85.6 | 97.5 | 98.9 | 577.7 |
| VL-BEIT [2] | 79.5 | - | - | 61.5 | - | - | - | 95.8 | - | - | 83.9 | - | - | - |
| $BLIP_{base}$ | 80.6 | 95.2 | 97.6 | 63.1 | 85.3 | 91.1 | 512.9 | 96.6 | 99.8 | 100.0 | 87.2 | 97.5 | 98.8 | 579.9 |
| **+Our method** | **81.4** | **95.7** | **98.1** | **63.6** | **85.9** | **91.8** | **516.5** | **97.3** | **100.0** | **100.0** | **87.9** | **97.9** | **99.1** | **582.2** |
| $OmniVL_{base}$ [30] | 82.1 | 95.9 | 98.1 | 64.8 | 86.1 | 91.6 | 518.6 | 97.3 | 99.9 | 100.0 | 87.9 | 97.8 | 99.1 | 582.0 |
| **+Our method** | **82.8** | **96.3** | **98.6** | **65.2** | **86.6** | **92.0** | **521.5** | **97.7** | 99.9 | **100.0** | **88.1** | **98.2** | **99.3** | **583.2** |
| $BLIP_{large}$ | 82.4 | 95.4 | 97.9 | 65.1 | 86.3 | 91.8 | 518.9 | 97.4 | 99.8 | 99.9 | 87.6 | 97.7 | 99.0 | 581.4 |
| **+Our method** | 83.2 | 95.9 | **98.4** | **65.6** | **87.0** | **92.7** | **522.8** | **98.0** | 99.9 | **100.0** | **88.7** | **98.0** | **99.2** | **583.8** |

mantic matching of intra-modal data and the inability of single-modal data to map to the same public space, limiting ITR model performance. We demonstrated that our method can be applied to any ITR retrieval model for universal retrieval. We applied our method to the datasets MSCOCO and flickr30k and combined it with the $BLIP_{base+ours}$ model, significantly improving their RSUM index. Specifically, on the MSCOCO dataset, our method improved the RSUM index of the $BLIP_{large+ours}$ model by 3.6%, while on the flickr30k dataset, the improvement was 2.4%. These results indicate that our proposed method can effectively enhance the performance of ITR models and has the potential for widespread applications in universal retrieval.

### 4.4. Zero Shot Retrieval

Zero-shot [10] retrieval is a method that utilizes information from unseen categories or samples in a search or retrieval task. In zero-shot retrieval, the model needs to reason on unseen categories or samples without being trained on them. Zero-shot retrieval can prompt the model to learn more generalized feature representations, thereby improving the model's generalization ability in unknown fields or categories. This helps improve the adaptability and generalization ability of the model, allowing it to be used in a wider range of application scenarios. We use the network fine-tuned on the MSCOCO data set to test the performance of zero-shot retrieval on Flickr30K, as shown in Table. 2, our method can learn the similarities and correlations between data in different fields and can effectively improve the model's generalization.

### 4.5. Effect of Different Parameters

In this subsection, we select representative hyperparameters $m$, $\alpha$, $\beta$ to explore the effect of different parameter combinations on the model performance. Where m is the number of layers of the MMLP, and  denote the weight coefficients of the two loss functions we optimise, respectively. We have also demonstrated through extensive experiments that the 3-layer MMLP structure can improve the retrieval results to the best. Figure. 6 shows the distribution of recall values obtained by MMLP with different number of layers for MSCOCO and Flickr30k datasets. With the Fig-

Table 2: Zero-shot image-text retrieval results on Flickr30K.

| Method | Flickr30k(1K test set) TR R@1 | TR R@5 | TR R@10 | IR R@1 | IR R@5 | IR R@10 |
|---|---|---|---|---|---|---|
| CLIP | 88.0 | 98.7 | 99.4 | 68.7 | 90.6 | 95.2 |
| ALIGN | 88.6 | 98.7 | 99.7 | 75.7 | 93.8 | 96.8 |
| ALBEF | 94.1 | 99.5 | 99.7 | 82.8 | 96.3 | 98.1 |
| $BLIP_{base}$ | 94.8 | 99.7 | 100.0 | 84.9 | 96.7 | 98.3 |
| **+our method** | **95.3** | **99.8** | **100.0** | **85.5** | **97.2** | **98.4** |
| $BLIP_{large}$ | 96.7 | 100.0 | 100.0 | 86.7 | 97.3 | 98.7 |
| **+our method** | **97.0** | 99.0 | **100.0** | **87.1** | **97.8** | **98.9** |

ure. 6, we can conclude that the performance of the model is optimal when the number of MMLP layers increases to 3. However, as the number of layers continues to increase to 4 and above, the performance starts to drop slightly, especially on the image retrieval task, with a tendency to fall back. In other words, increasing model complexity does not always lead to performance improvement, and overfitting or information redundancy occurs beyond a certain number of layers, resulting in a worse model. We use heat maps to show the effect of parameters ($\alpha$ and $\beta$) on the sum of R@K values for cross-modal retrieval RSUM. In the experiments for the dataset Flickr30K in Figure. 5a, the model achieves the highest RSUM 538.8 when the parameter combinations are $\alpha = 0.6$ and $\beta = 0.4$, which is clearly reflected in the heat map. It is observed that the RSUM shows a clear downward trend as the parameter combination is changed. For example, when $\alpha$ = 0.4 and $\beta$ = 0.8, the RSUM is only 534, which illustrates the importance of proper parameter configurations in terms of modal alignment and cross-modal alignment. This data suggests that adjacent parameter combinations (e.g., $\alpha$ = 0.6 and $\beta$ = 0.6) are equally capable of achieving relatively high performance, but fail to achieve optimality. Overall, the colour changes in the heatmap visually present the differences in model performance, further highlighting the impact of optimal parameter selection on RSUM. Through the Figure. 5b, we observe a significant effect of the modal alignment parameters ( and ) on the RSUM. When the parameters are set to $\alpha$ = 0.6 and $\beta$ = 0.4, the RSUM reaches a maximum value of 522.8, which clearly demonstrates the importance of the optimal parameters on the model performance. As the parameters are adjusted, the RSUM gradually decreases, especially for the combination of $\alpha$ = 1 and $\beta$ = 1, which has a RSUM of 516.9, showing a lower performance. The change of colours in the heatmap effectively reflects this phenomenon, with parameter combinations closer to the central value showing higher RSUM. This trend further confirms the necessity of proper parameter configuration in model training, and suggests that future research could explore more parameter ranges on this basis to seek further optimisation.

## 5. Ablation Study

We conducted an in-depth ablation study; the results are shown in the Table. 3. First, we study the role of the proposed method in the visual language model framework and perform ablation experiments on the single-modal momentum soft label alignment loss and the multi-modal data latent mapping method. The results show that regardless of removing the UMSA or MMLP method, the performance of the ITR retrieval model will be reduced. We use the BlIP model as the baseline model for our ablation experiments. BLIP is a pre-trained model that utilizes self-supervised and weakly supervised learning methods to unify visual and language understanding using large-scale image and text data. We used its $BLIP_{base}$ model as the benchmark and experimented with our proposed method. As shown in Table. 1, adding the method MMLP model to the benchmark model has a slight increase in performance compared to the benchmark model. This indicates that compared with the method of linearly mapping multimodal data into a shared space, MMLP can more effectively improve the adaptability and generalization of the model. Afterward, the UMSA method was added to the baseline

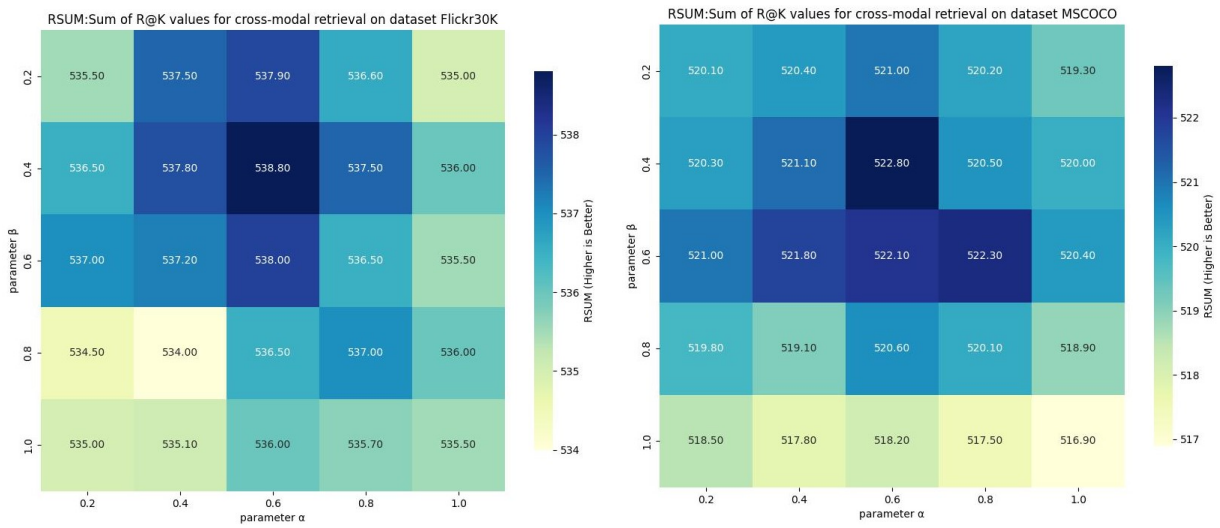
Table 3: Ablation study on fine-tuned image-text retrieval. The average recall on the test set is reported.

| Method | Flickr30k(1K test set) TR R@1 | TR R@5 | TR R@10 | IR R@1 | IR R@5 | IR R@10 |
|---|---|---|---|---|---|---|
| $BLIP_{base}$ | 96.6 | 99.8 | 100.0 | 87.2 | 97.5 | 98.8 |
| **+MMLP** | **96.9** | **99.9** | **100.0** | **87.6** | **97.8** | **99.0** |
| $BLIP_{base}$ | 96.6 | 99.8 | 100.0 | 87.2 | 97.5 | 98.8 |
| **+UMSA** | **97.1** | **99.9** | **100.0** | **87.8** | **98.2** | **99.3** |
| $BLIP_{base}$ | 96.6 | 99.8 | 100.0 | 87.2 | 97.5 | 98.8 |
| **+UMSA+MMLP** | **97.3** | **100.0** | **100.0** | **87.9** | **97.9** | **99.1** |

model for experiments. The results showed that UMSA improved model performance better than the baseline model. Both methods contributed to model performance improvement, but UMSA made a more significant contribution. We then added UMSA and MMLP to the benchmark model for experiments. We found that the experimental results were higher than those of adding the two methods separately to the model for training. Therefore, the proposed method can effectively improve the model's adaptability and generalization to unknown data through this experiment.

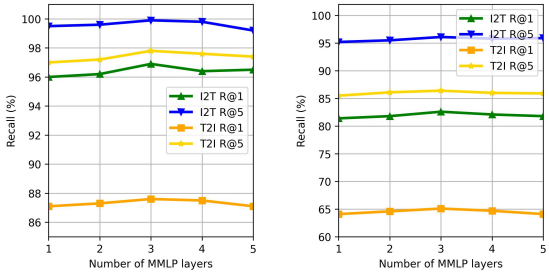
## 6. Case Study

The method proposed in this article and the $BLIP_{base}$ model were used to search for pictures and text in the same case, and the results were compared and analyzed. We compared the top three search results for images and text on the Flickr30K dataset. The green text and boxes represent content similar to the text and images, while the red text and




(a) (b)

Figure 5: Heatmap showing the RSUM for different combinations of and parameters. The central region represents the highest RSUM rate, with surrounding areas showing a gradual decrease as the parameters vary.




(a) Results on Flickr30K. (b) Results on MSCOCO.

Figure 6: Effect of the number of MMLP layers on the datasets of Flickr30k and MSCOCO, where I2T and T2I indicate text retrieval and image retrieval, respectively.

boxes indicate that they are not similar. From the retrieval results of Query a) on the benchmark model in Figure. 7, we can analyze that part of the data is lost (the 'grass' in the text is lost) due to the non-linear characteristics of the data not being considered, resulting in an incomplete match between the image and the text. From the retrieval results in Query b) in the Figure. 7, we can analyze that due to the lack of information interaction between text data, the final retrieval results are biased. In Figure. 8, we tested the model on the MSCOCO dataset, and we can observe that one of the benchmark models succeeds in retrieving the relevant image given the text in Query a), while the other two retrieve the content missing the objects 'horse' and 'carriages' respectively, and accordingly, when retrieving the relevant text given the image in Query b), the content retrieved by the benchmark model is partially contained in the image, although it is not the same as the text retrieved from the image. The result of the baseline model retrieved from the given image contains some of the content in the image, but some of the content does not exist in the image, so it is considered as a failed match. Our method takes both points into account and retrieves the content that best matches the input. The above analysis leads us to conclude that cross-modal alignment is certainly important, but intra-modal alignment and multi-modal data mapping distribution are also key to cross-modal retrieval.

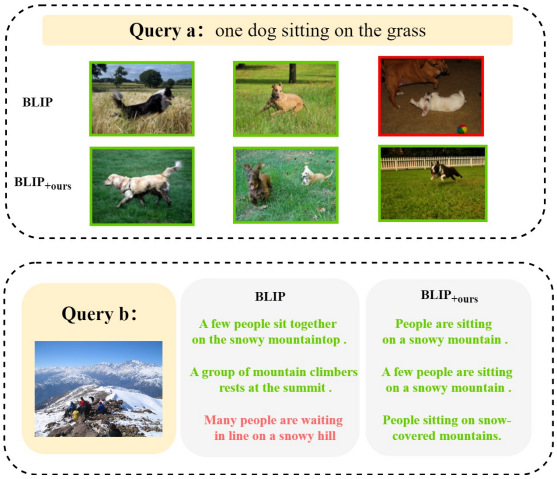



Figure 7: Comparison with the benchmark model BLIP retrieved on Flickr30K.

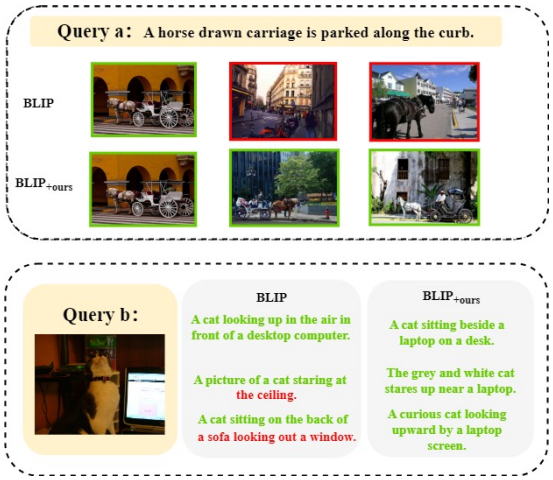



Figure 8: Comparison with the benchmark model BLIP retrieved on MSCOCO.

## 7. Conclusion

Our study is focused on the critical challenges of achieving alignment between vision and language in visual language training. To tackle these challenges, we propose a novel approach that simultaneously addresses the issues of single-modal data alignment and efficient mapping of different modal data into a unified common space. Our method is designed to provide soft-label supervision signals for the ITR (Image-Text Retrieval) model, relying on the unimodal pre-training model. Additionally, it uses the UMSA (Unimodal Momentum Soft Label Alignment) method to handle the multi-modal semantic alignment problem and strengthen the unimodal samples' similarity recognition.

Moreover, our MMLP (Multi-modal Data Latent Projection Method) method automatically adjusts the complexity of the model through non-linear modeling of complex data to better capture the data's characteristics and improve the model's expressive ability. To validate the effectiveness of our approach, we conducted extensive experimental verification covering various ITR models and datasets. The results show that our method significantly improves the performance of image-text retrieval.

It is worth highlighting that our method also enhances the generalization performance of the ITR model, improving its zero-shot performance on unseen data. These findings promote theoretical visual language research progress and provide innovative and effective solutions to visual-language tasks in practical applications. In summary, our proposed approach offers a promising way to achieve alignment between vision and language in visual language training. It has the potential to enhance the performance of ITR models significantly.

### Acknowledgments

This work was supported by the Shaanxi Province Science and Technology Innovation Leader Fund, China (No.TZ0336), supported by the Fundamental Research Funds for Central Universities, CHD, No.300102244301

## References

[1] H. Bao, W. Wang, L. Dong, Q. Liu, O. K. Mohammed, K. Aggarwal, S. Som, S. Piao, and F. Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912, 2022. 2, 9

[2] H. Bao, W. Wang, L. Dong, and F. Wei. Vl-beit: Generative vision-language pretraining. *arXiv preprint arXiv:2206.01127*, 2022. 9

[3] Y. Chen, Z. Ma, Z. Zhang, Z. Qi, C. Yuan, B. Li, J. Pu, Y. Shan, X. Qi, and W. Hu. How to make cross encoder a good teacher for efficient image-text retrieval? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26994–27003, June 2024. 9

[4] Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu. Uniter: Universal image-text representation learning, 2020. 8

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. 8

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 8

[7] E. Englesson and H. Azizpour. Generalized jensen-shannon divergence loss for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34:30284–30297, 2021. 5

[8] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017. 7

[9] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3

[10] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006. 9

[11] Z. Huang, Z. Zeng, B. Liu, D. Fu, and J. Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers, 2020. 3

[12] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 8, 9

[13] W. Kim, B. Son, and I. Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021. 2, 9

[14] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 201–216, 2018. 2, 9

[15] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2, 8

[16] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2

[17] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 9694–9705. Curran Associates, Inc., 2021. 8

[18] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020. 8

[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 7

[20] I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017. 8

[21] I. Loshchilov and F. Hutter. Decoupled weight decay regularization, 2019. 8

[22] L. Paninski. Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253, 2003. 6

[23] Z. Peng, L. Dong, H. Bao, Q. Ye, and F. Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers, 2022. 3

[24] X. Qin, L. Li, G. Pang, and F. Hao. Heterogeneous graph fusion network for cross-modal image-text retrieval. *Expert Systems with Applications*, 249:123842, 2024. 9

[25] X. Qin, L. Li, J. Tang, F. Hao, M. Ge, and G. Pang. Multi-task visual semantic embedding network for image-text retrieval. *Journal of Computer Science and Technology*, 39:811–826, 2024. 9

[26] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6, 8

[27] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge, 2015. 8

[28] H. Touvron, M. Cord, A. El-Nouby, J. Verbeek, and H. Jégou. Three things everyone should know about vision transformers, 2022. 8

[29] L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. 4

[30] J. Wang, D. Chen, Z. Wu, C. Luo, L. Zhou, Y. Zhao, Y. Xie, C. Liu, Y.-G. Jiang, and L. Yuan. Omnivl: One foundation model for image-language and video-language tasks. *Advances in neural information processing systems*, 35:5696–5710, 2022. 9

[31] Y. Wang, H. Yang, X. Qian, L. Ma, J. Lu, B. Li, and X. Fan. Position focused attention network for image-text matching. *arXiv preprint arXiv:1907.09748*, 2019. 2, 8

[32] Y. Wu, S. Wang, G. Song, and Q. Huang. Learning fragment self-attention embeddings for image-text matching. In *Proceedings of the 27th ACM international conference on multimedia*, pages 2088–2096, 2019. 2

[33] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 7

[34] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu. Coca: Contrastive captioners are image-text foundation models, 2022. 2, 3

[35] Y. Zeng, X. Zhang, H. Li, J. Wang, J. Zhang, and W. Zhou. X 2-vlm: All-in-one pre-trained model for vision-language tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[36] Y. Zeng, X. Zhang, H. Li, J. Wang, J. Zhang, and W. Zhou. $X^2$2-vlm: All-in-one pre-trained model for vision-language tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3156–3168, 2024. 2