IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 July 2024, accepted 20 August 2024, date of publication 22 August 2024, date of current version 4 September 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3447872*

RESEARCH ARTICLE

# MVItem: A Benchmark for Multi-View Cross-Modal Item Retrieval

**BO LI[1], JIANSHENG ZHU[2], LINLIN DAI[3], HUI JING[3], ZHIZHENG HUANG[3], AND YUTENG SUI[1]**

[1]Postgraduate Department, China Academy of Railway Sciences, Beijing 100081, China
[2]Department of Science, Technology and Information, China Railway, Beijing 100844, China
[3]Institute of Computing Technology, China Academy of Railway Sciences Corporation Ltd., Beijing 100081, China

Corresponding author: Jiansheng Zhu (zhujiansheng@sina.com)

This work was supported by the Foundation of China Academy of Railway Sciences under Grant 2023YJ138.

**ABSTRACT** The existing text-image pre-training models have demonstrated strong generalization capabilities, however, their performance of item retrieval in real-world scenarios still falls short of expectations. In order to optimize the performance of text-image pre-training model to retrieve items in real scenarios, we present a benchmark called MVItem for exploring multi-view item retrieval based on the open-source dataset MVImgNet. Firstly, we evenly sample items in MVImgNet to obtain 5 images from different views, and automatically annotate this images based on MiniGPT-4. Subsequently, through manual cleaning and comparison, we present a high-quality textual description for each sample. Then, in order to investigate the spatial misalignment problem of item retrieval in real-world scenarios and mitigate the impact of spatial misalignment on retrieval, we devise a multi-view feature fusion strategy and propose a cosine distance balancing method based on Sequential Least Squares Programming (SLSQP) to achieve the fusion of multiple view vectors, namely balancing cosine distance(BCD). On this basis, we select the representative state-of-the-art text-image pre-training retrieval models as baselines, and establish multiple test groups to explore the effectiveness of multi-view information on item retrieval to easing potential spatial misalignment. The experimental results show that the retrieval of fusing multi-view features is generally better than that of the baseline, indicating that multi-view feature fusion is helpful to alleviate the impact of spatial misalignment on item retrieval. Moreover, the proposed feature fusion, balancing cosine distance(BCD), is generally better than that of feature averaging, denoted as balancing euclidean distance(BED) in this work. At the results, we find that the fusion of multiple images with different views is more helpful for text-to-image (T2I) retrieval, and the fusion of a small number of images with large differences in views is more helpful for image-to-image (I2I) retrieval.

**INDEX TERMS** Cross-model retrieval, deep learning, item retrieval, contrastive text-image pre-training model, multi-view.

## I. INTRODUCTION

Item retrieval has extensive potential applications in fields such as lost item search and product search [1], [2], [3]. The objective of item retrieval is to find the item in a gallery that satisfy a given query, where the contents of the gallery are images and the query can be either an image or a textual description containing specific features. the current product datasets [2], [4], [5], [6] mainly consist of image-modal datasets with simple categorical labels, lacking effective textual description, which makes it difficult to apply method based this datasets to cross-modal retrieval task.

The contrastive text-image pre-trained cross-modal model, such as CLIP [7] and its derivatives [8], [9], [10], [11], due to its strong generalization capability, can serve as an off-the-shelf cross-modal item retrieval solution for achieving

The associate editor coordinating the review of this manuscript and approving it for publication was Gianluigi Ciocca.

rapid application deployment and saving development costs. However, applying these models to cross-modal item retrieval still faces challenges.

On the one hand, the prevailing image-text datasets, such as Flickr30k [12], MS COCO [13] and Laion-400m [14], contain a lot of content unrelated to item retrieval, such as scenery, people, and buildings. The test result of models on these datasets cannot be used to represent the performance of models on item retrieval.

On the other hand, in real-world scenarios, an image is a specific view sample of an item, while a query is typically a subjective and freely chosen description of the item from a specific view, provided by the user. We speculate

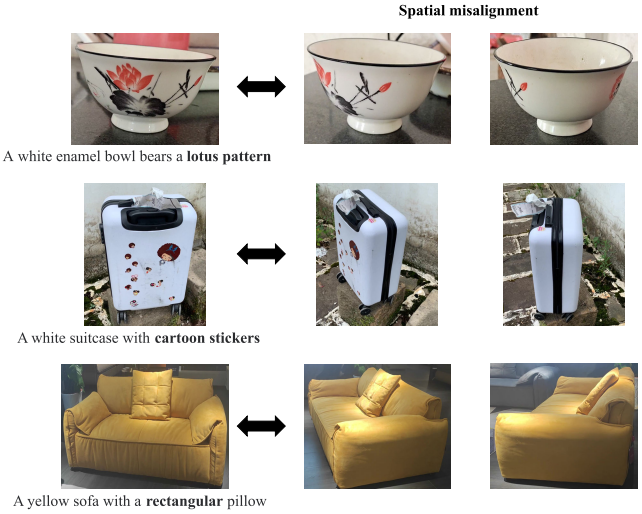


**FIGURE 1. Misalignment between the query and the positive images in the gallery.**

that if there is a spatial misalignment between the query descriptive information and the positive images in the gallery, as shown in Figure 1, the retrieval performance is inevitably affected. Therefore, we are concerned about the item retrieval capability of contrastive text-image pre-trained models in multi-view environment.

To address the aforementioned issue, we propose a test benchmark named MVItem to explore the item retrieval capability of contrastive text-image pre-trained retrieval models in multi-view setting. Specifically, we select 158 categories of items from the open-source multi-view item image dataset MVImgNet [15] and use MiniGPT-4 [16] for auxiliary image annotation, and finally get a high-quality text description for each item sample by manual cleaning and comparison. The content of these texts pertains to the color, shape, texture of items, as well as the relationship between items and background.

On basis of MVItem, we select the representative state-of-the-art text-image pre-training retrieval models to investigate the performance and explore rule of this models on item retrieval. Additionally, in order to solve the problem of spatial misalignment, we propose an item retrieval strategy based on multi-view feature fusion. Balancing euclidean distance is a common and effective feature fusion approach. However, we argue that uniformly integrating multi-view information is more beneficial for responding the subjectively random nature of real-world query scenarios. Therefore, we attempt to balance cosine distance, such that the fused feature vector is equiangular with each view feature vector. We obtain the image features of single-view by random sampling, and use cross-validation to simulate the retrieval performance of traditional retrieval methods as the baseline.

All in all, the contribution and key observations of our work can be summarized as follows.

- We propose a test benchmark for multi-view text-image item retrieval named MVItem, which is annotated by MiniGPT-4 based on open-source multi-view image dataset MVImgNet. Through manual cleaning and comparison, MVItem has 2376 high-quality text description for 11880 multi-view item images of 158 major categories.
- We test multiple representative the state-of-the-art text-image pre-training retrieval models, evaluate the performance of these models on item retrieval, analyze deficiency the impact of the design strategies of the models on item retrieval based on the experimental results.
- We devise a multi-view feature fusion strategy, design experimental baselines, and establish multiple test groups to explore the influence of multi-view information introduction on item retrieval which has potential spatial misalignment.
- We find balancing cosine distance (BCD) is generally better than that of balancing euclidean distance (BED) in multi-view image feature fusion. Meanwhile, the fusion of multiple images with different views is more helpful for text-to-image retrieval, and the fusion of a small number of images with large differences in views is more helpful for image-to-image retrieval.

## II. RELATED WORK

### A. ITEM SEARCH BASED ON IMAGE RETRIEVAL

Item search has a wide range of applications in the field of commodity retail, and its development relies on widely appearing commodity datasets, such as RPC [4], Grocery Products Dataset [17], RP2k [18], Supermarket Produce Dataset [19]. Image retrieval, as a primary technique for item search, usually involves encoding retrieval images into vectors using a pre-trained model of a categorization task [20]. The retrieval is accomplished by computing the similarity between vectors. Therefore, constructing discriminative feature vectors for image instances is crucial for image retrieval techniques, making contrastive learning the primary methodology for enhancing image retrieval performance, such as Siamese Loss [21], Triplet Loss [22] and InfoNCE [23].

### B. CROSS-MODAL RETRIEVAL

Cross-modal retrieval refers to the process of matching and searching between different modalities, aimed at eliminating

the heterogeneity differences among various modalities. This paper focus on investigating text-to-image retrieval. The primary approach for implementing text-to-image retrieval is to seek a method that maps text and image into a shared subspace [24], [25], [26], [27]. The emergence of vision-text pre-training model makes the model can be trained once and applied many times, which reduces the threshold of cross-modal retrieval application. CLIP [7] is a classic and representative cross-modal pre-training model. BLIP [28], [29] uses bootstrapping multitasking framework, and achieve the state-of-the-art results on text-to-image retrieval. The key advantage of pre-trained models lies in their zero-shot capability. Zero-shot capability is a highly sought-after performance in open category tasks [30], [31], [32], such as the item retrieval discussed in this paper.

### C. TEXT-TO-IMAGE RETRIEVAL BENCHMARK

Benchmark is a standard dataset that evaluates the performance of algorithms or models on a particular problem or task. Cola [33] is a benchmark for compositional text-to-image retrieval, which is used to explore empirical modeling designs to adapt pre-trained vision-text models to reason compositionally. T2I-VeRi [34] provides an evaluation platform for text-to-image vehicle re-identification. UFineBench [35] introduces a new benchmark for text-based person retrieval with ultra-fine granularity. HRS-Bench [36] provides a holistic and objective evaluation scheme for text-to-image synthesis tasks. PosterLayout [37] provides data and evaluation for content-aware visual-textual presentation layouts. According to the above research, The work of benchmark involves not only collecting and processing data relevant to specific problems, but also establishing reasonable evaluation metrics and designing appropriate experimental methodologies to explore the underlying issues and discover patterns.

## III. DESIGN PRINCIPLES OF MVItem BENCHMARK

### A. EMPHASIS ON ITEM RETRIEVAL TASK

Item retrieval is a special kind of open-category retrieval task. In contrast to general image retrieval, item retrieval does not include non-item samples such as animals, landscapes and buildings. Additionally, unlike specific image retrieval tasks like vehicle re-identification [34] and person retrieval [35], item retrieval tasks encompass a diverse range of categories that are not limited to a single major class. Therefore, the generalization advantage and zero-shot learning capability of contrastive text-image pre-training model make it more competent in the task of item retrieval. It is essential to establish a testing scenario that focuses on items in order to examine the performance of the model in item retrieval tasks.

### B. EXPLORING THE IMPACT OF SPATIAL MISALIGNMENT ON ITEM RETRIEVAL IN A MULTI-VIEW ENVIRONMENT

The current retrieval task usually ignores the multi-view factor. In real scenarios, user subjectively use image or description of an item from a certain view as the query:

$$Query = \left\{Img_{view_x}, Text_{view_x}\right\} \tag{1}$$

where $Img_{view_x}$ represents a query image for a certain view $x$, and $Text_{view_x}$ represents the corresponding query description for that view.

Gallery can be described as:

$$Gallery = \left\{\{N_1, \ldots, N_M\}, \left\{P_{view_y}\right\}\right\} \tag{2}$$

where $N_1$ to $N_M$ means there are $M$ negative samples, and $P_{view_y}$ means the positive sample a certain view $y$.

We are concerned about the presence of spatial misalignment between the view $x$ in query and view $y$ in gallery, which may affect the retrieval performance of the model.

Therefore, we need to devise a method to explore the influence of spatial misalignment on item retrieval in multi-view environment, heuristically find a strategy to alleviate spatial misalignment.

## IV. DATASET CONSTRUCTION

### A. SOURCE OF IMAGE DATA

Our work is based on MVImgNet [15], which is a large-scale dataset comprising multi-view images, which is originally a dataset for 3D reconstruction [38] and encompasses 6.5 million frames extracted from 219,188 videos encompassing objects belonging to 238 classes. The content of MVImgNet focuses on objects commonly encountered in daily-life, excluding scenery, animals, buildings, and other non-item entities, thus satisfying the requirement of image content in item retrieval.

MVImgNet provides labels for every major categories, but simple category labels cannot meet the needs of cross-modal item retrieval. Therefore, further descriptive text annotation work is needed.

### B. ANNOTATION AND CLEANING

#### 1) AUTOMATIC ANNOTATION BY MiniGPT-4

Relying solely on manual methods for descriptive annotation is highly subjective and inefficient. Therefore, we employ a lightweight large-scale vision-language generation model, MiniGPT-4 [16], to generate descriptive annotation information. MiniGPT-4 consists of a pretrained image encoder, a large language model (LLM), and a linear layer for aligning visual information to the LLM. The pretrained image encoder employed is a Vit transformer [8] combined with a Q-Former in BLIP-2 [29] and the LLM is Vicuna [39].

The annotation process is shown in Fig. 2. Each sample in MVImgNet consists of approximately 30 consecutive multi-view images, which exhibit minimal noticeable differences between adjacent images. Computing all of them would result in unnecessary computational overhead. Thus, we adopt an evenly sampling approach, capturing images from 5 distinct viewpoints. Next, we encode 5 images with different views respectively, and average the 5 image feature tensors to obtain the final <ImageHere> containing multi-view information.

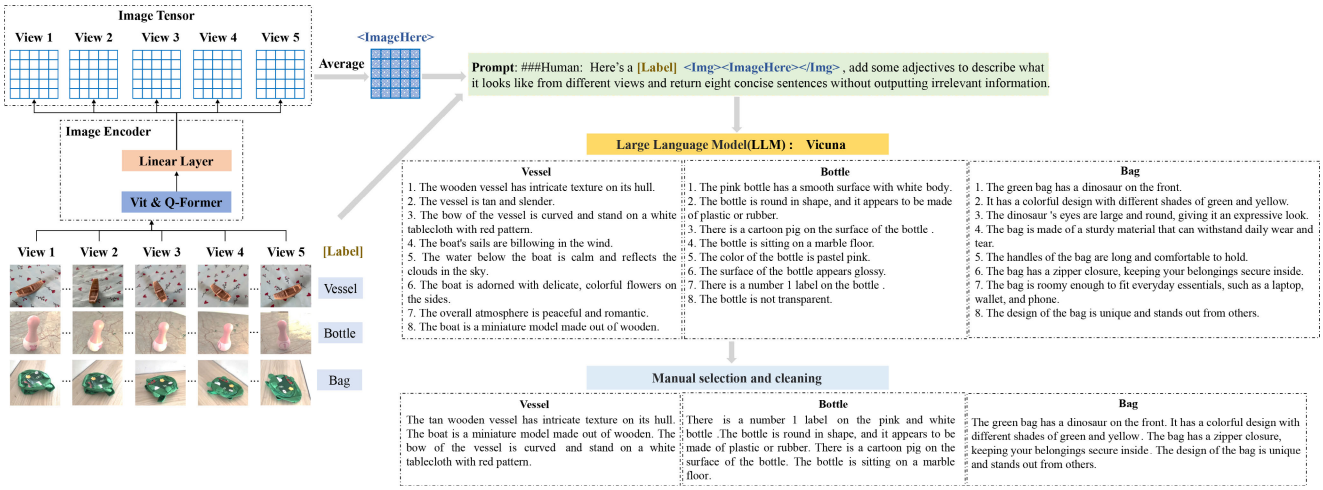


**FIGURE 2.** Annotation and cleaning workflow based on MiniGPT-4.


**FIGURE 3.** Image-text data in MVItem.

We have designed a prompt for annotation, incorporating the <ImageHere> and the sample label [Label], which is fed into Vicuna to generate 8 descriptive semantic annotations for manual selection:

**Prompt**: ###Human:Here's a **[Label]** <Img> **<ImageHere>**</Img>, add some adjectives to describe what it looks from different views and return eight concise sentences without outputting irrelevant information.### Assistant:**[descriptive annotation]**

We are conducting data annotation on a server equipped with four Tesla V100-DGXS-32GB GPUs running Ubuntu 20.04. The file management format follows MVImgNet, which ensures consistent major category numbers and sample IDs with MVImgNet.

### 2) MANUAL CLEANING AND COMPARISON

MVImgNet exists a significant amount of duplicate samples, necessitating manual de-duplication efforts in order to mitigate the issue of redundancy. In addition, MVImgNet includes some categories that are not suitable for item retrieval, such as scattered snacks, root of plant, etc. Therefore, we separate abnormal item categories and select the categories of artifact that are suitable for normal item retrieval need.

**TABLE 1. Major categories selected from MVImgNet.**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| bag | bottle | washer | toy vessel | toy train | telephone | table | stove | sofa | skatebord | rifle | pistol | remote control |
| printer | flowerpot | pillow | piano | mug | motorcycle | microwave | microphone | mailbox | loudspeaker | laptop | lamp | knife |
| pot | helmet | guitar | bookshelf | faucet | earphone | display | dishwasher | computer keyboard | clock | chair | car | cap |
| can | camera | cabinet | bus | bowl | bicycle | bench | bed | bathtub | basket | ashcan | airplane | umbrella |
| plush toy | toy figure | towel | toothbrush | toy bear | toy cat | toy bird | toy insect | toy cow | toy dog | toy monkey | toy elephant | toy fish |
| toy horse | toy sheep | toy mouse | toy tiger | toy rabbit | toy dragon | toy snake | toy chook | toy pig | rice cooker | pressure cooker | toaster | dryer |
| battery | curtain | blackboard eraser | bucket | calculator | candle | cassette | cup sleeve | computer mouse | easel | fan | coat rack | guitar stand |
| can opener | flashlight | hammer | scissors | screw driver | spanner | hanger | jug | fork | chopsticks | spoon | ladder | ceiling lamp |
| wall lamp | lamp post | light switch | mirror | paper box | wheelchair | walking stick | picture frame | shower | toliet | sink | power socket | bagged snacks |
| tripod | selfie stick | hair dryer | lipstick | glasses | sanitary napkin | tissue paper | rockery | book | projector | percussion | wind instruments | scarf |
| shoe | skirt | pants | clothing | box | soccer | power strip | toy turtle | bath sponge | glove | badminton | accessory | cigarette |
| stapler | lighter | key | watch | lock | hairbush | adhesive hook | bell | water pipe adapter | calendar | insecticide | electric saw | inflator |
| ironmongery | bulb | - | - | - | - | - | - | - | - | - | - | - |

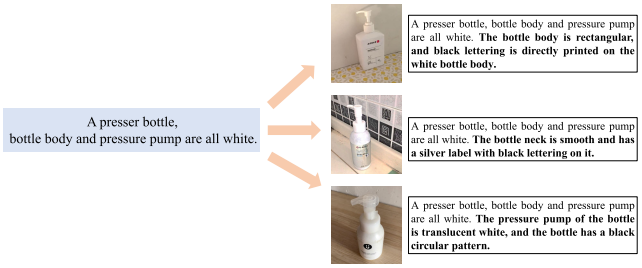


**FIGURE 4. Instance distinction.**

Moreover, It is worth noting that annotations based on MiniGPT-4 still show obvious hallucination. Therefore, we generated eight annotations using the model and then manually selected and cleaned 2-4 correct annotations, in order to eliminate the misrepresentation. Finally, these annotations are concatenated to form a complete description of the sample.

During the process of manual selection, it is crucial to ensure the completeness of the semantic information in the chosen annotations and avoid the occurrence of digressive and useless descriptions. Furthermore, efforts should be made to ensure that different annotations depict distinct content, thereby enriching the descriptive information from diverse perspectives and highlighting the features of each viewpoint.

Additionally, we allow the model to replace the label with synonyms, such as replacing ''vessel'' with ''boat'' or ''soccer'' with ''ball'' in order to preserve diversity in the descriptions.

The partial final results are shown in Fig. 3, which contains 5 images of a sample from different viewpoints and a complete sentence of description text.

It is worth noting that, in order to assess the reasonableness of evaluation, we need to ensure that each description corresponds to a single sample instance, avoiding situations where one description corresponds to multiple samples. Therefore, in the process of data cleansing, it is necessary to make further distinctions between approximate semantic annotations, as shown in the Fig. 4.

#### 3) DATA STATISTICS

We obtain 2376 non-duplicate item samples from 158 major categories of artifact sampled from MVImgNet, as shown in Tab.1, with a total of 11,880 images and 2376 text descriptions. Major category consists of multiple subcategories, in order to ensure the variety of samples. For instance, the bag category encompasses various types of bags such as backpacks, ladies shoulder bags, wallets, and briefcases.

## V. MULTI-VIEW FEATURE FUSION

### A. OVERVIEW

The spatial misalignment between query and the gallery can potentially affect the performance of retrieval. To address this issue, we propose a novel item retrieval strategy based on multi-view feature fusion to explore the influence of spatial misalignment on item retrieval in multi-view environment, as illustrated in Fig. 5.

Unlike traditional retrieval methods that use item image as the basic retrieval unit, we use a fused multi-view image feature of a sample as the fundamental retrieval unit in the gallery.

$$I_{gallery} = Fusion\left(I_{v_1}, I_{v_2}, \ldots, I_{v_n}\right) \tag{3}$$

where $I_{gallery}$ represents the multi-view image fusion feature of a sample in gallery, and $I_{v_1}, I_{v_2}, \ldots, I_{v_n}$ represent the image features of each sample view.

We aim to investigate the item retrieval performance of the state-of-the-art contrasting text-image pre-training models and design a baseline representing traditional retrieval way to verify the effectiveness of the multi-view feature fusion approach in alleviating spatial misalignment.

### B. FUSION STRATEGY

The essence of multi-view image feature fusion is to balance the distance between the query feature vector and the feature vectors of multi-view images of a sample, in order to introduce information from multiple views. Image multi-view features are fused before the actual query, and the query is unknowable at fusion operation. In the face of this situation, we have two approaches. One is to collect a large amount of query history information and design a feature fusion module to learn a fusion feature approach that conforms to the historical query preferences. Alternatively, in the absence of query participation, we can directly fuse the current multi-view image feature information.

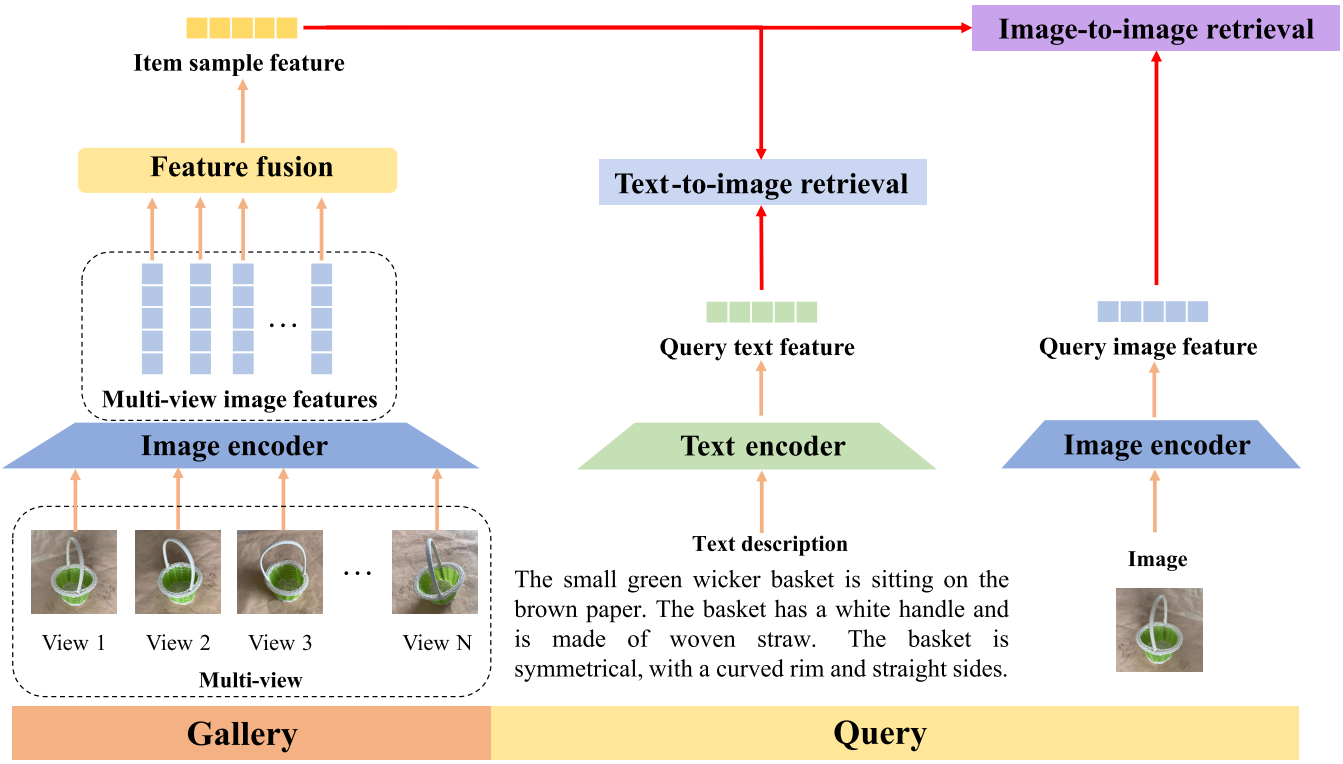


**FIGURE 5.** Item retrieval strategy based on multi-view feature fusion.

In this paper, as an exploratory approach, we directly fuse multi-view image feature information, and observe the performance of balancing euclidean distance and balancing cosine distance in item retrieval.

#### 1) BALANCING EUCLIDEAN DISTANCE(BED)

We can achieve balance in the euclidean distance by summing and averaging multi-view image feature vectors.

$$Fusion_{BED}\left(I_{v_1}, I_{v_2}, \ldots, I_{v_n}\right) = \frac{I_{v_1} + I_{v_2} + \ldots + I_{v_N}}{N} \quad (4)$$

where $N$ represents the number of views.

#### 2) BALANCING COSINE DISTANCE(BCD)

The vector fusion method of BED tends to bias the final fused vector towards some specific given vectors. However, it is desirable to obtain a fusion result that is not skewed towards any given vector. Balanced cosine distance means to find a vector such that the angle between the vector and the given multi-view image feature vector is equal. In the above problems, there are generally multiple vectors satisfying the conditions, and we need to obtain the vector with the smallest angle with the multi-view image feature vector as the fusion feature.

In cases where the dimensionality of vectors is high, obtaining exact solutions using linear algebra methods can be challenging. Therefore, we resort to optimization algorithms, Sequential Least Squares Programming (SLSQP), to obtain the optimal solution.

##### *a: OBJECTIVE FUNCTION*

We choose the sum of the squared differences in angles between the target vector and the multi-view image feature vectors as the objective function.

$$f\left(I_x\right) = \sum_{i=1}^{N}\left(\frac{I_x.I_{v_i}}{||I_x||_2.||I_{v_i}||_2} - \sum_{j=1}^{N}\frac{I_x.I_{v_j}}{||I_x||_2.||I_{v_j}||_2}/N\right)^2 \quad (5)$$

where $I_x$ represents the target vector for the iterative calculation, and $||\cdot||_2$ represents the magnitude of the vector.

##### *b: CONSTRAINT CONDITION*

The image features of N views of a sample are similar, and we want to obtain an average vector that is approximately consistent with the feature vectors of these N views. Therefore, the angle between the target vector and the image feature vectors of N views should be smaller than the maximum angle among these N viewpoint vectors.

$$\forall\, n \in N, g\left(I_x\right) = cosine(I_x, I_{v_n}) \quad (6)$$

$$g\left(I_x\right) \leqslant \max_{k,w \in N} cosine\left(I_{v_k}, I_{v_w}\right) \quad (7)$$

##### *c: LAGRANGE FUNCTION FOR ITERATIVE OPTIMIZATION*

The Lagrangian function for iterative training is constructed based on $f\left(I_x\right)$ and $g\left(I_x\right)$.

$$L\left(I_x, \lambda\right) = f\left(I_x\right) + \lambda * g\left(I_x\right) \quad (8)$$

**TABLE 2. Multi-view image feature cosine distance balancing algorithm based on Sequential Least Squares Programming (SLSQP).**

| **Multi-view image feature cosine distance balancing algorithm** |
|---|
| **Input**: $f(I_x)$, $g(I_x)$, $I_x$, error tolerance(tol). |
| **Output**:Cosine distance balancing vector $I_{BCD}$. |
| **Initialization**: $I_x$, set iteration counter k=1, calculate the gradient of $f(I_x)$ and $g(I_x)$ |
| **While** the tol is not met **do**: |
| 1. Construct and solve the Lagrange function $L(I_x, \lambda)$, and get updating direction $d_k$. |
| 2. Determined update step $a_k$ by backtrack search, and update $I_x = I_x + a_k * d_k$ . |
| 3. Evaluate $g(I_x)$, update $f(I_x)$ and $g(I_x)$, $k = k + 1$, calculate the gradient of $f(I_x)$ and $g(I_x)$ . |
| $I_{BCD} = I_x$ |
| **Return** $I_{BCD}$ |

where $\lambda$ represents Lagrange coefficient. We choose error tolerance as the stopping criterion for iteration, and the overall algorithmic workflow is illustrated in Tab. 2.

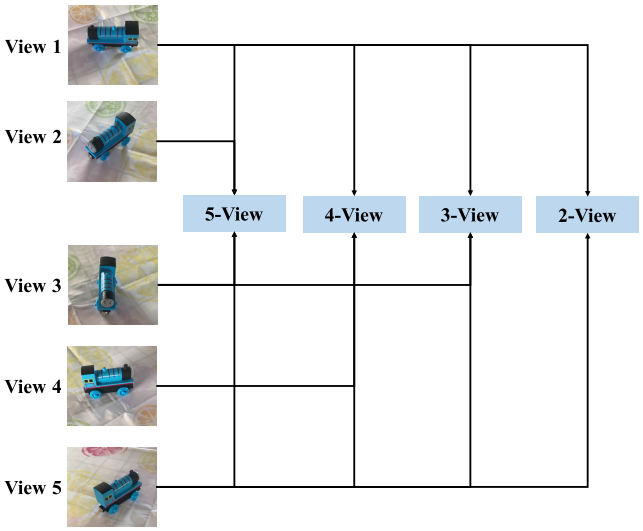


**FIGURE 6. Multi-view feature fusion test groups.**

## VI. BENCHMARK SETTING

### A. SETTING

#### 1) BASELINE DESIGN

We assume that adopting the fusion feature of multi-view images in gallery can alleviate the impact of spatial misalignment on item retrieval. In this case, the retrieval unit in the gallery is a fusion feature of the sample multi-view image, and there is only one positive sample in the gallery.

Correspondingly, our baseline group uses single-view images in gallery as the retrieval unit, and there is also only one positive sample. We adopt cross-validation to ensure the randomness of positive sample in the gallery. Each validation group randomly selects one view image from each sample as a retrieval unit in the gallery. We severally set 50 cross-verification groups for Image-to-image(I2I) and Text-to-image(T2I) retrieval tests, and took the average results as the baseline for I2I and T2I.

#### 2) THE INFLUENCE OF MULTI-VIEW FEATURE FUSION

By controlling the fused multi-view image number of item sample, we explore the influence of multi-view information introduction on item retrieval which has potential spatial misalignment. Specifically, we set up the fusion of 2-5 images with the biggest view difference of a sample to observe the T2I performance effect between them and the baseline respectively, as shown Fig. 6.

Similarly, we also have established multi-view fusion test group for I2I as well. For each MVItem sample, one image from the set of five images is selected as the query, and the remaining 2-4 images are fused based on maximizing the difference in viewpoints, resulting in the creation of four test groups.

### B. MODELS SELECTED FOR EVALUATION

We are not only concerned with the performance of current CLIP-related the state-of-the-art pre-trained text-image models on item retrieval tasks, but also aim to further elucidate the impact of the underlying design strategies of these models on item retrieval. Therefore, we summarize the technical characteristics of the selected model:

#### 1) CLIP [7]

The fundamental idea of CLIP is to create an $N \times N$ matrix of image-text pairs, while simultaneously training an image encoder and a text encoder, with the aim of maximizing the cosine similarity of $N$ positive image-text pairs and minimizing the cosine similarity of $N^2 - N$ negative text pairs. CLIP uses cross-entropy loss function (InfoNCE-like) to achieve contrastive representation learning for multi-class N-pair.

#### 2) LONG-CLIP [40]

Vanilla CLIP integrates absolute positional encoding with a suggested text input length limit of 77 tokens, which makes it challenging to apply in scenarios that require detailed descriptions. Therefore, Long-CLIP extends the text input length to 248 characters and performs fine-tuning on long-text image-text pairs to enhance the model's fine-grained retrieval capabilities.

#### 3) ALBEF [41]

The distinctive feature of ALBEF lies in aligning the image-text features before the fusion and incorporating momentum distillation. Feature alignment also adopts Image-Text contrastive Learning similar to CLIP to achieve semantic alignment between images and text at the overall level. Image-Text Matching and Masked text Modeling are used to complete the training of feature fusion. ALBEF introduces Momentum Distillation, where Momentum model is the exponential moving average of ALBEF that is constantly trained. Momentum Distillation makes each training of the

model take into account the prediction distribution of the previous version.

#### 4) CoCa [9]

The distinctive feature of CoCa is to pre-train the encoder by adding generative pre-trained task branches of captioning. CoCa can be approximated as adding an Encoder-Decoder Captioning to CLIP. Contrastive learning still employs the cross-entropy loss function and utilizes the maximum likelihood function to quantify the self-regressive captioning loss.

#### 5) SLIP [10]

The distinctive feature of SLIP is to combine the visual representation contrast learning of SimCLR [42] with the image-text contrast learning of CLIP. SLIP simultaneously emphasizes the role of image information in the model while aligning modalities.

#### 6) BLIP-2 [29]

BLIP-2 leverages a frozen image encoder to bootstrap text-image representation learning and achieves multimodal semantic alignment through Image-Text contrastive learning similar to CLIP. The bootstrapping process involves a image-grounded text generation task and multiple trainable queries(Usually 32). Q-former is the main content of BLIP-2. We mainly pay attention to the performance of image encoder and text encoder of Q-former on item retrieval.

### C. RETRIEVAL PERFORMANCE METRIC

#### 1) RANK@1

Accuracy is a crucial metric for evaluating retrieval models. In the context of multi-view feature fusion, where the gallery contains only one feature vector representing the positive sample, it is essential to assess the model's ability to retrieve this positive sample in one query. We use $Rank@1$ to represent the accuracy of a single retrieval of the model.

$$Rank@1 = \frac{P}{N} \times 100\% \tag{9}$$

where $N$ indicates the total number of queries, and $P$ indicates the correct number of queries.

#### 2) MEAN SIMILARITY DISTRIBUTION (mSD)

In addition to accuracy, we also need to quantitatively evaluate the model's feature extraction and alignment effect. Mean similarity distribution (mSD) [35] deems continuous similarity values more realistically reflect the model's retrieval ability, thus overcoming the challenge faced by mean Average Precision (mAP) in accurately quantifying a model's retrieval performance from the same ranking results.

Item retrieval requires the ability to quantitatively assess the model with precision, rather than simply evaluating its performance based on continuous similarity pattern. Therefore, we choose mSD as the metric to evaluate the performance level of item retrieval.

Let $S^+$ represents positive sample, $S^-$ represents negative sample, the first $n$ return results of item retrieval be $Top_n$:

$$Top_n = \{S_1, S_2, \ldots, S_n\} \tag{10}$$

mSD requires computing the positive samples and negative samples ratio (PNR) in $Top_n$ as well as the average similarity precision (ASP).

PNR calculate the normalized average similarity ratio between positive and negative:

$$PNR = 1 - e^{-kx} \tag{11}$$

where $x$ is the average similarity ratio between positive and negative samples, and $k$ is set to 1 as default.

ASP can be calculated as:

$$ASP = \frac{1}{n^+} \sum_{k=1}^{n^+} \frac{\sum_{i=1}^{j_k} s_i^+}{\sum_{i=1}^{j_k} s_i} \tag{12}$$

where $\{j_1, j_2, \ldots, j_{n^+}\}$ represents the rankings of positive samples.

The similarity distribution (SD) of $Top_n$ can be calculated by the product of PNR and ASP. Finally, mSD can be obtained by averaging SDs of all rank lists.

## VII. EXPERIMENTS

### A. IMPLEMENTATION DETAILS

We conduct experiments based on MVItem to investigate the performance of different pre-trained models on item retrieval. Specifically, we explore the impact of multi-view feature fusion on addressing potential spatial misalignment in item retrieval from both text-to-image and image-to-image perspectives.

The pre-trained models actually come in multiple versions. In order to fairly explore the performance of different models on item retrieval and analyze the impact of model design strategies on item retrieval. We uniformly select the model version under the Vit-L/14 image encoder. Specifically, CoCo [43] and Flickr30k [12] are currently the two most frequently used types of multimodal general-purpose datasets. To compare the performance of these two datasets on the item retrieval task, we select the open-source models ALBEF_CoCo and ALBEF_Flickr30k. In the process of evolution, we utilize a plug-and-play text encoder based on Long-CLIP to achieve text feature encoding for texts exceeding 77 tokens of CLIP. We calculate mean similarity distribution(mSD) using top-5 returned results. The experiments are performed on 1 GeForce RTX 4090.

### B. EXPERIMENTAL RESULT

#### 1) TEXT-TO-IMAGE RETRIEVAL (T2I)

The results of the text-to-image item retrieval are shown in Tab. 3 and Tab. 4.

##### a: ITEM RETRIEVAL PERFORMANCE OF PRE-TRAINED MODELS

The baseline reflects the fundamental capabilities of the model, with performance rankings on the T2I task as follows:

**TABLE 3.** Rank@1 of text-to-image item retrieval in MVItem.

| Model | BED | | | | BCD | | | | Baseline |
|---|---|---|---|---|---|---|---|---|---|
| | 2-view | 3-view | 4-view | 5-view | 2-view | 3-view | 4-view | 5-view | |
| CLIP [7] | 37.23% | 38.16% | **38.35%** | **38.35%** | 37.23% | 38.19% | **38.35%** | **38.35%** | 35.41% |
| Long−CLIP [40] | 39.73% | 40.26% | **40.32%** | **40.32%** | 39.73% | 40.26% | **40.32%** | 40.30% | 36.66% |
| ALBEF_CoCo [41] | 21.75% | 22.33% | 22.55% | 22.12% | 21.75% | 22.33% | **22.63%** | 22.13% | 21.11% |
| ALBEF_Flickr30k [41] | 24.24% | 24.72% | 24.77% | 24.75% | 24.24% | 24.73% | **24.81%** | 24.75% | 21.91% |
| CoCa [9] | 58.87% | **60.91%** | 60.13% | 60.26% | 59.27% | 60.72% | 60.19% | 60.31% | 56.07% |
| SLIP_Large [10] | 12.59% | 12.51% | 12.75% | 12.77% | 12.63% | 12.51% | 12.77% | **12.80%** | 11.05% |
| BLIP2_CoCo [29] | 50.13% | 49.61% | 50.07% | 50.29% | 50.36% | 50.09% | 50.15% | **51.13%** | 46.26% |

**TABLE 4.** Top-5 mSD of text-to-image item retrieval in MVItem.

| Model | BED | | | | BCD | | | | Baseline |
|---|---|---|---|---|---|---|---|---|---|
| | 2-view | 3-view | 4-view | 5-view | 2-view | 3-view | 4-view | 5-view | |
| CLIP [7] | 29.8718 | 30.5987 | 30.9306 | 31.0374 | 29.9013 | 30.662 | 30.9781 | **31.0781** | 28.7812 |
| Long−CLIP [40] | 33.8084 | 34.2275 | 34.3954 | 34.4064 | 33.9304 | 34.2332 | **34.4081** | 34.3498 | 31.4896 |
| ALBEF_CoCo [41] | 23.6539 | 23.9102 | 23.9556 | 23.8382 | 23.7106 | 23.9371 | **24.0113** | 23.8721 | 21.6893 |
| ALBEF_Flickr30k [41] | 23.6412 | 23.7286 | 24.0838 | 23.8524 | 23.6891 | 23.8235 | **24.1036** | 23.8729 | 21.7038 |
| CoCa [9] | 44.7037 | **45.4201** | 45.1279 | 45.2534 | 44.7929 | 45.4189 | 45.1254 | 45.2591 | 42.8573 |
| SLIP_Large [10] | 14.4595 | 14.4565 | 14.6376 | 14.6591 | 14.4602 | 14.4592 | 14.6569 | **14.6631** | 12.7255 |
| BLIP2_CoCo [29] | 40.0116 | 39.8231 | 39.9742 | 40.0813 | 40.0921 | 39.9802 | 40.0213 | **40.1293** | 37.8517 |

CoCa > BLIP2 > LongCLIP > CLIP > ALBEF_Flickr30k > ALBEF_CoCo > SLIP_Large.

The commonality between CoCa and BLIP2 lies in the presence of a cross attention module in their model structures, distinguishing them from subsequent models lacking this particular architecture. This suggests that cross attention plays a significant role in cross-modal alignment.

The superior performance of Long-CLIP over CLIP suggests that models trained with fine-grained text descriptions exhibit improved cross-modal alignment. Therefore, for item retrieval task, it is advisable to collect a more extensive set of detailed item description texts in order to enhance the model's capability to adapt to nuanced descriptions.

For ALBEF, compared with CoCo, Flickr30k has a more detailed text description, which makes the model trained based on Flickr30k better than the model trained based on CoCo on T2I when the model structure is the same. At the same time, even ALBEF also has cross-attention modules, the difference in the dimensions of training data scale and diversity between datasets, like Flickr30k and CoCo, and the large-scale training data collected from the Internet, like CLIP, has led to significant disparities between ALBEF and CLIP in the performance of item retrieval tasks based on MVItem.

It is worth noting that SLIP performs poorly on T2I, indicating that the direct introduction of image self-supervised learning on the basis of cross-modal contrast learning may significantly weaken the effect of cross-modal alignment.

### *b: ANALYSIS OF PERFORMANCE CHARACTERISTICS ON ITEM RETRIEVAL*

Tab. 3 illustrates that the current state-of-the-art text-image pre-training models still struggle to achieve perfect performance in multi-view item retrieval. To summarize and analyze the performance characteristics of models in item retrieval, we conduct tests specifically focusing on the bag category and present retrieval negative examples for each model in a single query, as shown in Fig. 7.

We can find that as the performance of the retrieval model improves, the characteristics of negative examples undergo change. In our experiment, SLIP performs the worst in T2I, with the negative example having only vague correspondences between certain words and the query. For example, in Fig. 7, it might have wrongly retrieved Mickey Mouse's bag for "cartoon," even though in the positive case, "cartoon" refers to a chick.

With the improvement in performance, some accurate correspondence between negative examples and queries can be achieved. Such as in the first retrieval example of ALBEF_Flickr30k, the negative case accurately corresponds to the word "golden". However, due to the neglect of descriptions like "chain straps", it incorrectly retrieves a handbag decorated with a golden sheen.

In terms of characteristics, CLIP appears to be a performance boundary, with negative examples being globally matched with the queries. The added descriptions have some impact on retrieval, however, there still exists a significant difference between negative examples and positive examples.

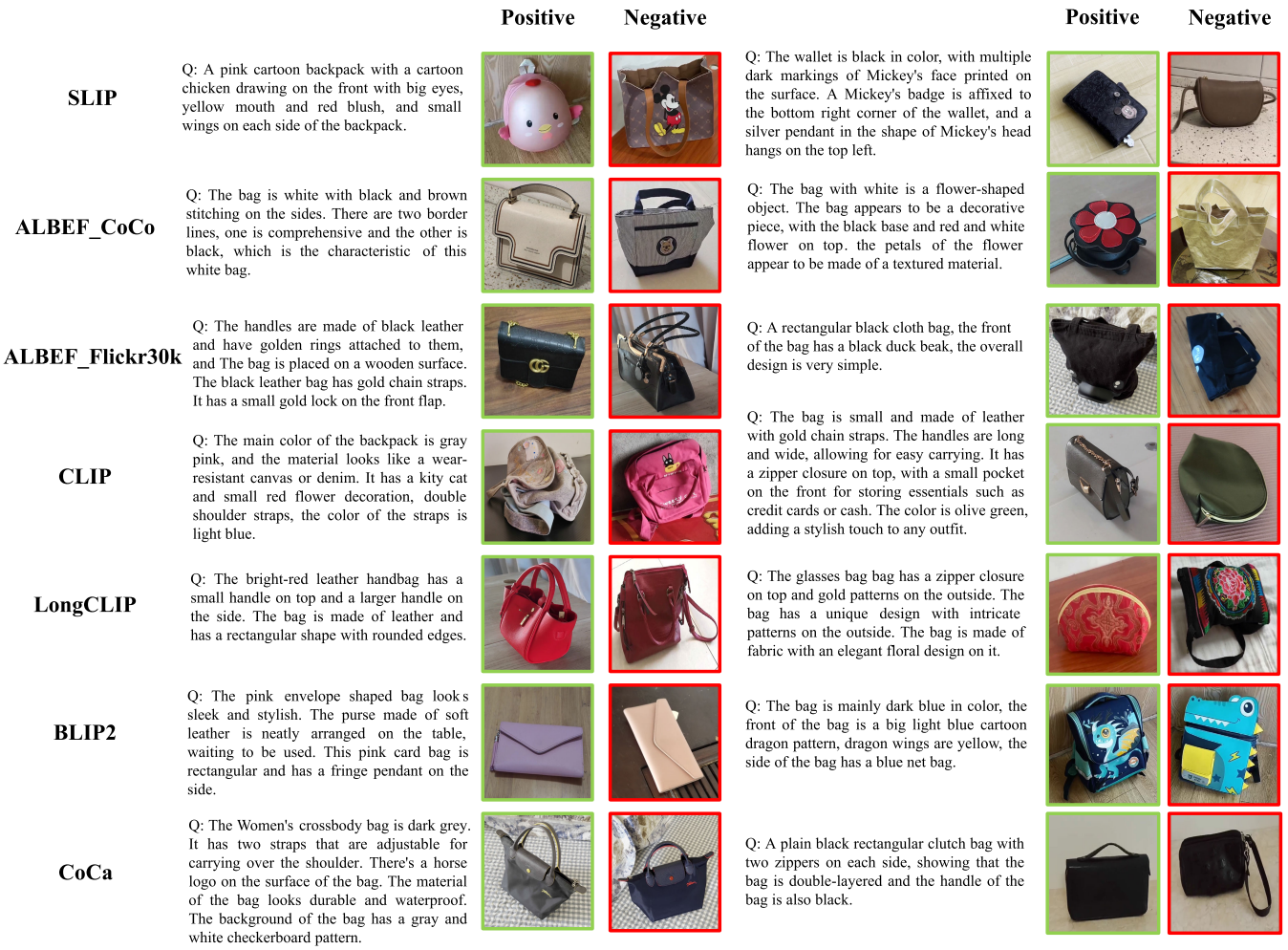


**FIGURE 7.** Partial retrieval negative examples in bag category.

Models with better performance compared to CLIP exhibit a high degree of similarity between the negative and positive examples regarding their types and characteristics, with errors typically arising from more detailed and specific descriptive attributes. For instance, in the first retrieval example of BLIP-2, the negative example is also an envelope-shaped bag, differing from the positive example in terms of color depth and the absence of a fringe pendant. As seen in CoCa, the negative and positive examples are nearly identical in characteristics, with errors often stemming from the query itself being insufficiently detailed or ambiguous.

### c: INFLUENCE OF SPATIAL MISALIGNMENT ON T2I

The superior retrieval performance of the multi-view feature fusion over the baseline indicates the non-negligible impact of spatial misalignment on T2I in real-world scenarios. Tab. 4 provides a more detailed reflection of the performance difference between post-feature fusion and pre-feature fusion, as shown by the mSD values. From Tab. 4, it can be observed that the overall performance of BCD is superior to that of BED, even though there a instance where the performance of BED rarely exceeds that of BCD, such as with CoCa where

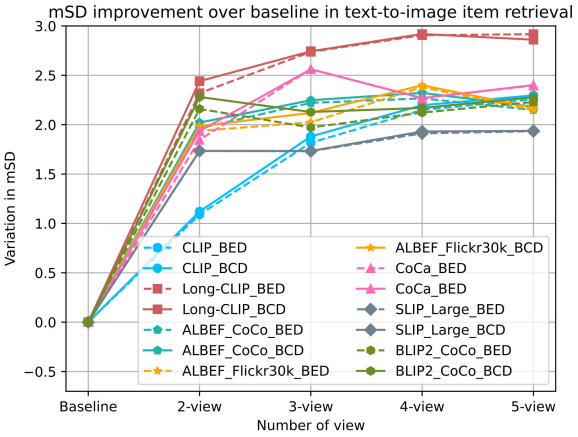


**FIGURE 8.** The relationship between the number of views and mean similarity distribution (mSD) variation in text-to-image item retrieval.

the BCD value is 45.4189 for the 3-view case, very closed to 45.4201 for BED.

The introduction of noise in multi-view images is inevitable, as it may include images with the greatest

**TABLE 5.** Rank@1 of image-to-image item retrieval in MVItem.

| Model | BED | | | BCD | | | Baseline |
|---|---|---|---|---|---|---|---|
| | 2-view | 3-view | 4-view | 2-view | 3-view | 4-view | |
| CLIP [7] | **98.63%** | 98.60% | 98.60% | **98.63%** | 98.60% | 98.60% | 85.81% |
| Long−CLIP [40] | **99.73%** | 99.68% | 99.68% | **99.73%** | 99.68% | 99.70% | 93.73% |
| ALBEF_CoCo [41] | 97.12% | 97.12% | 97.12% | 97.12% | **97.36%** | 97.12% | 90.93% |
| ALBEF_Flickr30k [41] | 98.11% | 98.11% | 98.11% | 98.11% | **98.37%** | 98.11% | 90.89% |
| CoCa [9] | **99.46%** | 99.35% | 99.35% | **99.46%** | 99.35% | 99.35% | 91.97% |
| SLIP_Large [10] | **99.83%** | 99.79% | 99.79% | **99.83%** | 99.79% | 99.79% | 94.61% |
| BLIP2_CoCo [29] | **98.91%** | 98.83% | 98.83% | **98.91%** | 98.83% | 98.83% | 93.12% |

**TABLE 6.** Top-5 mSD of image-to-image item retrieval in MVItem.

| Model | BED | | | BCD | | | Baseline |
|---|---|---|---|---|---|---|---|
| | 2-view | 3-view | 4-view | 2-view | 3-view | 4-view | |
| CLIP [7] | 63.3857 | 63.3378 | 63.3143 | **63.3902** | 63.3373 | 63.3139 | 57.7569 |
| Long−CLIP [40] | **64.3104** | 64.0589 | 64.0566 | 64.3096 | 64.0859 | 64.1608 | 60.8277 |
| ALBEF_CoCo [41] | 61.8078 | 61.8176 | 61.7943 | 61.8101 | **61.8237** | 61.8004 | 59.9821 |
| ALBEF_Flickr30k [41] | 62.0873 | 61.9582 | 61.9511 | 62.0881 | **62.1021** | 61.9875 | 59.1411 |
| CoCa [9] | 63.9029 | 63.7902 | 63.6432 | **63.9125** | 63.7562 | 63.6481 | 60.5324 |
| SLIP_Large [10] | **64.5468** | 64.4893 | 64.4257 | 64.5301 | 64.4691 | 64.4986 | 62.1412 |
| BLIP2_CoCo [29] | 63.2641 | 63.0896 | 63.0041 | **63.3781** | 63.1075 | 63.0618 | 60.5997 |

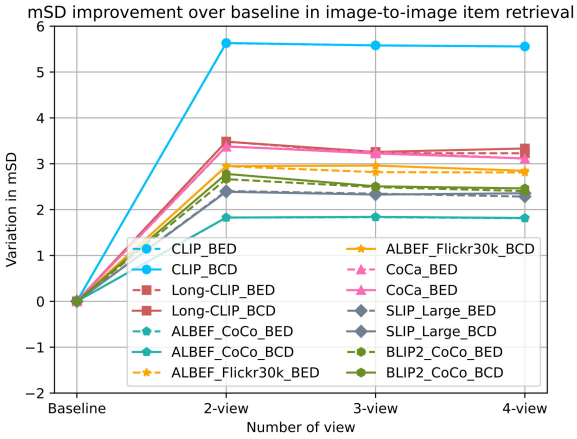


**FIGURE 9.** The relationship between the number of views and mean similarity distribution (mSD) variation in image-to-image item retrieval.

differences in viewing angles compared to the query. This noise can potentially lead to slightly lower performance in experimental groups that integrate a greater number of views compared to those with fewer views. However, this does not detract from the advantages that multi-view image fusion brings to T2I applications.

We have summarized the variation in mSD over the baseline on T2I after multi-view feature fusion, as shown in the Fig. 8. With the increase in the number of views, the mSD shows an increasing trend, indicating that the fusion of multiple images of one item with different views is beneficial for improving their T2I performance.

### 2) IMAGE-TO-IMAGE RETRIEVAL (I2I)

In addition, we also explore image-to-image retrieval of text-image pre-training models. The results of the image-to-image item retrieval are shown in Tab. 5 and Tab. 6.

#### *a: PERFORMANCE OF PRE-TRAINED MODELS*

The performance of the model in the I2I task is significantly better than that in the T2I task, with the performance ranking under baseline as follows: SLIP_Large > LongCLIP > BLIP2 > CoCa > ALBEF_CoCo > ALBEF_Flickr30k > CLIP.

SLIP demonstrates outstanding performance on I2I tasks, indicating that incorporating self-supervised image learning on the basis of cross-modal contrast learning can significantly enhance the discriminative power of the model for image features.

LongCLIP demonstrates superior performance indicating that fine-grained textual descriptions are advantageous for enhancing the understanding and discriminative capabilities of models for image. Conversely, CLIP exhibits a comparatively inferior performance on the baseline, suggesting that short texts and ambiguous descriptions may have an impact on the image feature construction for item.

#### *b: INFLUENCE OF SPATIAL MISALIGNMENT ON I2I*

The superior retrieval performance of the multi-view feature fusion over the baseline also indicates the non-negligible impact of spatial misalignment on I2I in real-world scenarios. It is worth noting that, unlike T2I, the fusion of multi-view features leads to significant changes in the ranking

of the model's I2I capability compared to the baseline. Specifically, from the perspective of mSD, SLIP_Large > LongCLIP > CoCa > CLIP > BLIP2 > ALBEF_Flickr30k > ALBEF_CoCo.

We consider the fusion performance of BCD on I2I to be generally superior to that of BED. Even though rare results of BCD exhibits slightly lower scores compared to that of BED on Long-CLIP (64.3096 for BCD vs. 64.3104 for BED) and CoCa (64.5301 for BCD vs. 64.5468 for BED). We guess that the subtle differences in performance may be attributed to variances between optimization and precision calculations.

We have also summarized the variation in mSD over the baseline on I2I after multi-view feature fusion, as shown in the Fig. 9. We find that CLIP exhibits the most significant improvement in I2I tasks after multi-view fusion. With the increase in the number of views, the variation in mSD is relatively flat but shows a subtle downward trend, indicating that the fusion of a small number of images with significant view disparities plays a more beneficial role in I2I.

## VIII. CONCLUSION AND DISCUSSION

In this paper, we have constructed a multi-view text-image item dataset to explore the performance of state-of-the-art text-image pre-training retrieval models in simulated real-world item retrieval scenarios. We collect multiple views of item samples from the open-source multi-view item dataset MVImgNet, and design an annotation process for multi-view images based on MiniGPT-4. Through manual cleaning and comparison, we obtain a test set for multi-view item retrieval, which consists of 2376 samples belonging to 158 major categories, each accompanied by a high-quality description and images from five different views.

To investigate the potential impact of spatial misalignment in item retrieval under real-world scenarios, we propose an item retrieval strategy based on multi-view feature fusion. Building upon this approach, we select the current state-of-the-art text-image pre-training retrieval models, design experimental baselines, establish performance evaluation metrics, create multiple test groups, and conduct experiments to explore the performance. Our main findings are as follows:

- Potential spatial misalignment phenomenon in real world scenarios has a significant impact on item retrieval, but this effect can be effectively alleviated through multi-view image feature fusion. In the context of text-to-image retrieval, the fusion of multiple images with different views may be more effective. However, for image-to-image retrieval, a small number of images with large differences in views may be more effective.
- In the multi-view image feature fusion process, balancing cosine distance (BCD) is more effective than balancing euclidean distance (BED).
- The model incorporating cross attention modules, when provided with a sufficiently large training data set size comparable to that of CLIP, exhibits strong performance in text-to-image item retrieval tasks.
- On the basis of contrastive learning like CLIP, directly incorporating image self-supervised learning significantly decreases the performance of text-to-image retrieval, while significantly boosting the performance of image-to-image retrieval.

In summary, based on MVItem, we explore the performance of text-image pre-trained cross-modal models on item retrieval in real-world scenarios and design effective methods and strategies to mitigate the spatial misalignment issue. In the future, it is necessary to build a large-scale multi-view text-image training dataset, which could help to promote the development of item retrieval.

## REFERENCES

[1] X. Dong, X. Zhan, Y. Wei, X. Wei, Y. Wang, M. Lu, X. Cao, and X. Liang, "Entity-graph enhanced cross-modal pretraining for instance-level product retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 13117–13133, Nov. 2023.

[2] X. Zhan, Y. Wu, X. Dong, Y. Wei, M. Lu, Y. Zhang, H. Xu, and X. Liang, "Product1M: Towards weakly supervised instance-level product retrieval via cross-modal pretraining," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11782–11791.

[3] P. Rahayu, D. I. Sensuse, B. Purwandari, I. Budi, F. Khalid, and N. Zulkarnaim, "A systematic review of recommender system for e-Portfolio domain," in *Proc. 5th Int. Conf. Inf. Educ. Technol.*, Jan. 2017, pp. 21–26.

[4] X.-S. Wei, Q. Cui, L. Yang, P. Wang, and L. Liu, "RPC: A large-scale retail product checkout dataset," 2019, *arXiv:1901.07249*.

[5] K. Georgiadis, G. Kordopatis-Zilos, F. Kalaganis, P. Migkotzidis, E. Chatzilari, V. Panakidou, K. Pantouvakis, S. Tortopidis, S. Papadopoulos, and S. Nikolopoulos, "Products-6K: A large-scale groceries product recognition dataset," in *Proc. 14th Pervasive Technol. Rel. Assistive Environ. Conf.*, 2021, pp. 1–7.

[6] Y. Bai, Y. Chen, W. Yu, L. Wang, and W. Zhang, "Products-10K: A large-scale product recognition dataset," 2020, *arXiv:2008.10545*.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[8] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, and Y. Cao, "EVA: Exploring the limits of masked visual representation learning at scale," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 19358–19369.

[9] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "CoCa: Contrastive captioners are image-text foundation models," 2022, *arXiv:2205.01917*.

[10] N. Mu, A. Kirillov, D. Wagner, and S. Xie, "SLIP: Self-supervision meets language-image pre-training," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 529–544.

[11] Y. Chen, X. Qi, J. Wang, and L. Zhang, "DisCo-CLIP: A distributed contrastive loss for memory efficient CLIP training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 22648–22657.

[12] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2641–2649.

[13] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. Lawrence Zitnick, "Microsoft COCO captions: Data collection and evaluation server," 2015, *arXiv:1504.00325*.

[14] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki, "LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs," 2021, *arXiv:2111.02114*.

[15] X. Yu, M. Xu, Y. Zhang, H. Liu, C. Ye, Y. Wu, Z. Yan, C. Zhu, Z. Xiong, T. Liang, G. Chen, S. Cui, and X. Han, "MVImgNet: A large-scale dataset of multi-view images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 9150–9161.

[16] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "MiniGPT-4: Enhancing vision-language understanding with advanced large language models," 2023, *arXiv:2304.10592*.

[17] M. George and C. Floerkemeier, "Recognizing products: A per-exemplar multi-label image classification approach," in *Proc. 13th Eur. Conf. Comput. Vis. (ECCV)*. Zurich, Switzerland: Springer, 2014, pp. 440–455.

[18] J. Peng, C. Xiao, and Y. Li, "RP2K: A large-scale retail product dataset for fine-grained image classification," 2020, *arXiv:2006.12634*.

[19] A. Rocha, D. C. Hauagge, J. Wainer, and S. Goldenstein, "Automatic fruit and vegetable classification from images," *Comput. Electron. Agricult.*, vol. 70, no. 1, pp. 96–104, Jan. 2010.

[20] W. Chen, Y. Liu, W. Wang, E. M. Bakker, T. Georgiou, P. Fieguth, L. Liu, and M. S. Lew, "Deep learning for instance retrieval: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7270–7292, Jun. 2023.

[21] K. Musgrave, S. Belongie, and S.-N. Lim, "A metric learning reality check," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*. Glasgow, U.K.: Springer, 2020, pp. 681–699.

[22] J. Song, T. He, L. Gao, X. Xu, and H. Tao Shen, "Deep region hashing for efficient large-scale instance search from images," 2017, *arXiv:1701.07901*.

[23] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9729–9738.

[24] Y. Liu, Y. Guo, E. M. Bakker, and M. S. Lew, "Learning a recurrent residual fusion network for multimodal matching," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4107–4116.

[25] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5005–5013.

[26] L. Ma, Z. Lu, L. Shang, and H. Li, "Multimodal convolutional neural networks for matching image and sentence," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2623–2631.

[27] Y. Huang, Q. Wu, C. Song, and L. Wang, "Learning semantic concepts and order for image and sentence matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6163–6171.

[28] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 12888–12900.

[29] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 19730–19742.

[30] K. You, J. Gu, J. Ham, B. Park, J. Kim, E. K. Hong, W. Baek, and B. Roh, "CXR-CLIP: Toward large scale chest X-ray language-image pre-training," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2023, pp. 101–111.

[31] H. Ma, H. Zhao, Z. Lin, A. Kale, Z. Wang, T. Yu, J. Gu, S. Choudhary, and X. Xie, "EI-CLIP: Entity-aware interventional contrastive learning for e-commerce cross-modal retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18051–18061.

[32] C. Bustos, C. Civit, B. Du, A. Solé-Ribalta, and A. Lapedriza, "On the use of vision-language models for visual sentiment analysis: A study on CLIP," in *Proc. 11th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2023, pp. 1–8.

[33] A. Ray, F. Radenovic, A. Dubey, B. Plummer, R. Krishna, and K. Saenko, "Cola: A benchmark for compositional text-to-image retrieval," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024, pp. 1–13.

[34] L. Ding, L. Liu, Y. Huang, C. Li, C. Zhang, W. Wang, and L. Wang, "Text-to-image vehicle re-identification: Multi-scale multi-view cross-modal alignment network and a unified benchmark," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 7, pp. 7673–7686, Jul. 2024.

[35] J. Zuo, H. Zhou, Y. Nie, F. Zhang, T. Guo, N. Sang, Y. Wang, and C. Gao, "UFineBench: Towards text-based person retrieval with ultra-fine granularity," 2023, *arXiv:2312.03441*.

[36] E. M. Bakr, P. Sun, X. Shen, F. F. Khan, L. E. Li, and M. Elhoseiny, "HRS-bench: Holistic, reliable and scalable benchmark for text-to-image models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 20041–20053.

[37] H. Y. Hsu, X. He, Y. Peng, H. Kong, and Q. Zhang, "Posterlayout: A new benchmark and approach for content-aware visual-textual presentation layout," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 6018–6026.

[38] Q. Wang, Z. Wang, K. Genova, P. Srinivasan, H. Zhou, J. T. Barron, R. Martin-Brualla, N. Snavely, and T. Funkhouser, "IBRNet: Learning multi-view image-based rendering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4690–4699.

[39] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing, "Vicuna: An open-source chatbot impressing GPT-4 with 90%* chatgpt quality," Mar. 2023. [Online]. Available: https://lmsys.org/blog/2023-03-30-vicuna/

[40] B. Zhang, P. Zhang, X. Dong, Y. Zang, and J. Wang, "Long-CLIP: Unlocking the long-text capability of CLIP," 2024, *arXiv:2403.15378*.

[41] 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 *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 9694–9705.

[42] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[43] 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 *Proc. 13th Eur. Conf. Comput. Vis. (ECCV)*. Zurich, Switzerland: Springer, 2014, pp. 740–755.

**BO LI** received the M.S. degree from Chongqing University, Chongqing, China, in 2022. He is currently pursuing the Ph.D. degree in traffic information engineering and control with China Academy of Railway Sciences, Beijing, China.

He is currently researching and exploring the integration and application of artificial intelligence technology in the railway passenger transportation scenarios, with a special focus on the use of vision-language pre-training models to find lost items in railway passenger transportation. His current research interests include computer vision and deep learning, specifically, including cross-modal retrieval, object detection, pre-training model, and self-supervised learning. He was awarded the title of "Outstanding Graduate Student" from Chongqing University, in 2020.

**JIANSHENG ZHU** received the Ph.D. degree in traffic information engineering and control from China Academy of Railway Sciences, Beijing, in 2002.

He is currently a Researcher and a Ph.D. Supervisor with China Academy of Railway Sciences. He has been a leading figure in China National "Thousand Talents Plan" and the Railway "Hundred Thousand Million Talents" Project. He has long been engaged in the informatization construction of railway passenger transportation and the key technological research of the ticketing system, responsible for the research on railway intelligence and big data applications, and has achieved multiple innovative research results and theoretical breakthroughs.

**LINLIN DAI** received the M.S. degree in network and switching technology from Beijing University of Posts and Telecommunications, Beijing, China, in 2007.

From 2007 to 2010, she was a Software Research and Development Engineer and the Project Manager of the IBM China System and Technology Development Center. Since 2010, she has been a Software Research and Development Engineer. She is currently the Technical Director of the Fundamental Platform Department, Institute of Computing Technology, China Academy of Railway Sciences. She has been working on system formulation, development promotion, and scientific research in the fields of infrastructure platform construction, intelligent service, software, and hardware development for passenger transport.

**HUI JING** received the M.S. degree in electronic and communication engineering from Beijing Jiaotong University, Beijing, China, in 2018.

Since 2018, he has been an Artificial Intelligence Algorithm Engineer with the Institute of Fundamental Platform Department, Institute of Computing Technology, China Academy of Railway. He has been deeply engaged in the research of artificial intelligence technology in railway passenger transportation scenarios, achieving breakthroughs in key technologies for facial recognition in railway passenger transportation, and innovative applications. His technological achievements have been widely applied throughout China railway system. His current research interests include facial recognition, speech recognition, and large-scale models.

**ZHIZHENG HUANG** received the M.S. degree in probability theory and mathematical statistics from the Renmin University of China, Beijing, China, in 2017.

From 2017 to 2023, he has devoted himself to the research of artificial intelligence algorithms and technology applications. As a Key Member, he has successfully completed several commercial AI technology projects, including the Bingo Smart Cashier and intelligent ordering robots. Since 2023, he has been an Artificial Intelligence Algorithm Engineer with the Institute of Fundamental Platform Department, Institute of Computing Technology, China Academy of Railway, and mainly responsible for the research and development of intelligent document recognition systems.

**YUTENG SUI** received the M.S. degree from Qingdao University of Science and Technology, Qingdao, Shandong, China, in 2019. He is currently pursuing the Ph.D. degree in traffic information engineering and control with China Academy of Railway Sciences, Beijing, China.

From 2019 to 2023, he was an Artificial Intelligence Algorithm Engineer with the Institute of Fundamental Platform Department, Institute of Computing Technology, China Academy of Railway. He is primarily responsible for the key technical research and product application of facial recognition turnstile gates in railways. His current research interests include facial recognition and 3D reconstruction.