

Rebalanced Vision-Language Retrieval Considering Structure-Aware Distillation

Yang Yang, *Member, IEEE*, Wenjuan Xi, *Student Member, IEEE*, Luping Zhou, *Senior Member, IEEE*, Jinhui Tang, *Senior Member, IEEE*

Abstract—Vision-language retrieval aims to search for similar instances in one modality based on queries from another modality. The primary objective is to learn cross-modal matching representations in a latent common space. Actually, the assumption underlying cross-modal matching is modal balance, where each modality contains sufficient information to represent the others. However, noise interference and modality insufficiency often lead to modal imbalance, making it a common phenomenon in practice. The impact of imbalance on retrieval performance remains an open question. In this paper, we first demonstrate that ultimate cross-modal matching is generally sub-optimal for cross-modal retrieval when imbalanced modalities exist. The structure of instances in the common space is inherently influenced when facing imbalanced modalities, posing a challenge to cross-modal similarity measurement. To address this issue, we emphasize the importance of meaningful structure-preserved matching. Accordingly, we propose a simple yet effective method to rebalance cross-modal matching by learning structure-preserved matching representations. Specifically, we design a novel multi-granularity cross-modal matching that incorporates structure-aware distillation alongside the cross-modal matching loss. While the cross-modal matching loss constraints instance-level matching, the structure-aware distillation further regularizes the geometric consistency between learned matching representations and intra-modal representations through the developed relational matching. Extensive experiments on different datasets affirm the superior cross-modal retrieval performance of our approach, simultaneously enhancing single-modal retrieval capabilities compared to the baseline models.

Index Terms—Vision-Language Retrieval, Imbalanced Multi-Modal Learning, Structure-Aware Distillation.

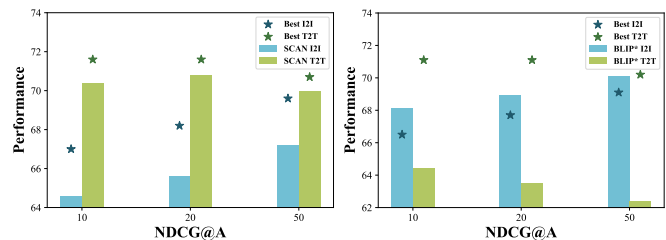
I. INTRODUCTION

IN re-examining current vision-language retrieval tasks [1, 2, 3], approaches typically focus on searching images for given texts or retrieving texts from image queries [4].

Manuscript received March 16, 2024; revised September 30, 2024 and November 12, 2024; accepted December 3, 2024. Date of current version December 12, 2024. This work was supported in part by the National Key RD Program of China (2022YFF0712100), in part by the NSFC (62276131), in part by the Natural Science Foundation of Jiangsu Province of China under Grant (BK20240081, BG2024042), and in part by the Fundamental Research Funds for the Central Universities (No.30922010317). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sicheng Zhao. (*Corresponding author: Jinhui Tang.*)

Yang Yang, Wenjuan Xi and Jinhui Tang are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: yyang@njust.edu.cn; xiwenjuan@njust.edu.cn; jinhuitang@njust.edu.cn). Luping Zhou is with the School of Electrical and Information Engineering, The University of Sydney, Sydney, NSW 2006, Australia (e-mail: luping.zhou@sydney.edu.au).

This paper has supplementary downloadable material available at <http://ieeexplore.ieee.org>, provided by the author. The material includes supplemental material. Contact yyang@njust.edu.cn for further questions about this work.



(a). Both modalities are simultaneously adversely affected. (b). “Weak” modality slightly improves, “strong” modality seriously declined.

Fig. 1. The impact of imbalanced modalities on single-modal retrieval in cross-modal learning. Figures (a) and (b) respectively present the $NDCG@\{10,20,50\}$ results of the single-modal encoders after training two cross-modal models, SCAN [6] and BLIP [14] on the MS-COCO (1K) dataset. The asterisk (*) indicates that the large models were retrained from scratch. “Best” denotes the results obtained by the best single-modal models trained separately on images and texts. The cross-modal model and single-modal models adopt identical network architectures.

Unlike single-modal retrieval, the primary challenge in vision-language retrieval lies in the semantic divergence of heterogeneous data, necessitating effective constraints to ensure the consistency between two modal representations [5].

To solve this problem, state-of-the-art (SOTA) approaches [6, 7] usually adopt various modal interaction modules to govern the matching of aligned modal representations for each instance, i.e., pulling cross-modal representations closer while distancing them from other instances. Initial methods are dual-encoder approaches [1, 6, 8, 9], in which the two modalities interact at the **output level**. These methods typically build independent embedding networks for each modality, and then design coarse-grained (e.g., global-level [8]) or fine-grained (e.g., region-level [6, 10, 11] and graph-level [12]) similarity functions to measure the matching degree of cross-modal output representations. With the development of transformer-based architectures, vision-language transformers are proposed [13], in which the two modalities interact from the **input level**. These methods adopt the deep transformer as a modal interaction module to collectively model the concatenation of two modal inputs, which can allow cross-modal instances to represent each other effectively.

Cross-modal matching representation learning [15] can be regarded as a variant of unsupervised single-modal contrastive learning, with the aligned modal instances serving as the positive anchors. However, experiments suggest that the effectiveness of aligned modal instances is inferior to the data augmentation technique in single-modal contrastive learning. This

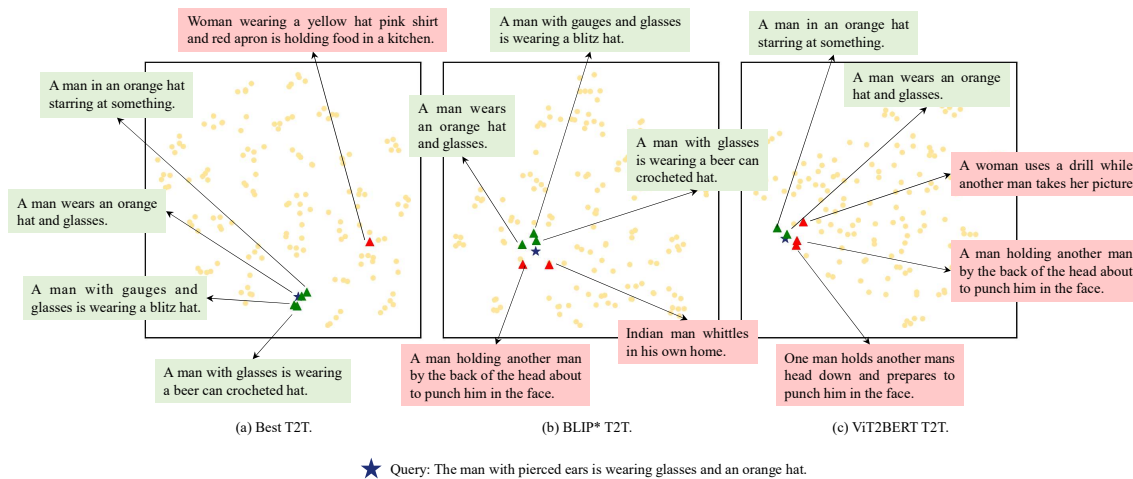


Fig. 2. T-SNE visualization of Best T2T, BLIP* T2T, ViT2BERT T2T on the FLICKR30K dataset, where BLIP* T2T uses the ViT/B and BERT as backbones. We randomly choose a text query and a database with 150 samples. The blue pentagram represents the text query, while the top-5 retrieval candidates are shown as triangles. Ground-truth candidates are marked in green, and non-ground-truth candidates are marked in red.

discrepancy arises from the noise interference and modality insufficiency in real applications [16, 17], leading to notorious modality imbalance phenomenon [18, 19]. Here, the modality with greater sufficiency is defined as strong modality [19], while the less sufficient one is termed weak modality. In real-world applications, we can employ the performance of the optimal single-modal model to assess the “strength” of a modality [16]. We use an experiment for illustration in Fig. 1. From Fig. 1, we observe that the text-to-text (T2T) retrieval performance of the text model BERT [20] consistently outperformed the image-to-image (I2I) retrieval performance of the vision model Swin Transformer (SwinT) [21], designating text as the “strong” modality and image as the “weak” modality. Furthermore, a simultaneous decrease in the expressive capabilities of both the weak and strong modalities, as shown in Fig. 1(a), and a slight improvement in the performance of the weak modality, accompanied by a severe decline in the performance of the strong modality, as shown in Fig. 1(b). This suggests that enforcing cross-modal consistency learning disrupts the optimal instance structure within single-modal spaces, thereby diminishing the accuracy of similarity-based retrieval. This issue is further evident in the t-SNE [22] visualization in Fig. 2, where the structural information in the text of BLIP* and ViT2Bert is partially compromised compared with the best text retrieval model BERT.

To address the modal imbalance problem, coined as the information gap between the two modalities, we propose a shift from exact instance-level modal matching to the structure preservation of the learned representations in cross-modal representation learning. In particular, we design a multi-granularity distillation module to rebalance cross-modal matching, which improves instance-level matching through representation-level and structure-aware distillation to construct better latent structures. The newly introduced geometric consistency can promote geometric symmetry in the latent common space. Specifically, for inter- and intra-modal structure preservation, we introduce two modal-independent teacher networks, which can jointly distill the cross-modal model

using the designed relational matching. Moreover, to obtain the optimal instances’ relationships from teacher models, we adaptively learn the fusion coefficient of two teachers. In summary, the main contributions of this paper are as follows: 1). We analyze the impact of the exact matching on cross-modal retrieval tasks. We show that reducing the modal gap at the instance-level does not always ensure better latent structure particularly when dealing with imbalanced modalities. Instead, a large information gap between the modalities can hurt the performance. 2). We advocate preserving both semantic and structural consistency, and propose the multi-granularity distillation on top of the cross-modal consistency loss, enhancing multi-granularity matching in latent common space. 3). Through extensive experiments, our approach shows improved performance across cross-modal, single-modal, and mixed retrieval. This validates the effectiveness of incorporating relational knowledge in the learning of comprehensive matching representations.

II. RELATED WORK

A. Vision-Language Retrieval

To learn matching representations of heterogeneous modalities, a large number of vision-language retrieval models are proposed [23, 24, 25, 26]. Traditional approaches [1, 6, 8, 9, 27] always utilize dual architecture, where the image and text modalities are separately embedded into a common space and then maximize the cross-modal representation similarity. For example, [8] built two independent modal encoders (i.e., VGG19 for image and GRU for text), and incorporated hard negatives in the triplet loss function; [6] further utilized the Faster R-CNN for image modality, and discovered the full latent alignments using both image regions and words in a sentence as context; CLIP [28] employed a large-scale contrastive pre-training approach to effectively align vision and language modalities; [29] applied contrastive learning to cross-modal hashing using a momentum-based optimizer and a cross-modal ranking learning loss. These approaches can efficiently

TABLE I
SINGLE-MODAL RETRIEVAL PERFORMANCE COMPARISON. EVALUATION CRITERIA IS NDCG@A (@A FOR SIMPLICITY).

Methods	MS-COCO (1K)						MS-COCO (5K)						FLICKR30K						Vizwiz					
	I2I			T2T			I2I			T2T			I2I			T2T			I2I			T2T		
	@10	@20	@50	@10	@20	@50	@10	@20	@50	@10	@20	@50	@10	@20	@50	@10	@20	@50	@10	@20	@50	@10	@20	@50
ResNet101	51.9	54.3	58.4	-	-	-	42.9	44.9	48.3	-	-	-	49.0	51.6	56.2	-	-	-	49.4	52.1	56.5	-	-	-
SwinT	66.5	67.7	69.1	-	-	-	61.5	63.2	65.7	-	-	-	64.1	64.5	67.9	-	-	-	50.3	53.9	57.3	-	-	-
LSTM	-	-	-	63.6	65.5	66.9	-	-	-	58.9	61.2	63.7	-	-	-	57.9	60.4	62.1	-	-	-	51.2	54.0	57.4
BERT	-	-	-	71.1	71.1	70.2	-	-	-	63.4	64.8	66.1	-	-	-	71.6	70.2	68.4	-	-	-	55.5	57.5	59.6

TABLE II
COMPARISON WITH SINGLE-MODAL RETRIEVAL MODELS (SMR), THE DISTILLATION IMAGE MODEL TAUGHT BY TEXT MODEL (S2W@IMAGE), AND THE DISTILLATION TEXT MODEL TAUGHT BY IMAGE MODEL (W2S@TEXT).

Method	I2I			T2T		
	@10	@20	@50	@10	@20	@50
SMR	64.1	64.5	67.9	71.6	70.2	68.4
S2W@Image	64.2	65.6	68.3	-	-	-
W2S@Text	-	-	-	71.5	70.0	68.0

benefit from numerous simple modal interactions such as dot products or shallow attention layers [10]. In contrast, inspired by the significant advances in language understanding led by Transformers [20], SOTA retrieval models turn to employ large vision-language transformers for modal interaction [1, 30]. For example, [30] developed a large-scale cross-modal encoder with five diverse representative pre-training tasks; [1] processed both visual and textual inputs that interact through co-attention transformer layers; [31] introduced a contrastive loss to align the image and text representations before fusing through cross-modal attention. In these approaches, vision and language inputs are fed into a unified or separate cross-modal attention branch to compute the similarity between them, thereby obtaining more similar cross-modal representations. However, we experimentally observe that the learned matching representations from cross-modal retrieval approaches will be affected when existing imbalanced modalities.

B. Imbalanced Multi-Modal Learning

The important assumption behind cross-modal methods is the modal balance, i.e., it typically assumes that each modality can well represent other modalities [17]. However, in real-world scenarios, factors such as data noise multi-modal data [16], heterogeneity of multi-modal data and missing multi-modal data can lead to modality imbalance. For instance, the noise can impact the model's ability to learn effective information, resulting in the sufficiency of different modalities is various, leading to the existence of strong and weak modalities [18, 19]. In traditional cross-modal matching constraints, strong and weak modalities will interact with each other, and it is difficult to control the strong modal structure from being affected by the weak modality when exiting a large cross-modal divergence. This problem leads to multi-modal joint training being hard and has been researched in multi-modal fusion for classification tasks, [32, 33, 34] observed that the best single-modal network often outperforms the multi-modal networks. Therefore, [32] proposed to estimate the single-

modal generalization and overfitting speeds to calibrate the learning through loss re-weighting; [33] promoted reliability and robustness by integrating evidence that explained the prediction of each modality; [34] chose to slow down the learning rate of the mighty modality by online modulation to lessen the inhibitory effect on the other modality. Nevertheless, these methods only focus on multi-modal fusion tasks by learning complementary information, which is different from the cross-modal retrieval for learning matching representations.

III. PROPOSED METHOD

To rebalance the vision-language retrieval, we aim to transfer the optimal instances' structure from teacher networks to cross-modal model in learning matching representations. Therefore, the overall architecture consists of two networks: 1). a cross-modal student network, and 2). two-modal independent teacher networks. In this paper, we concentrate on the image and text modalities, considering the effectiveness and feasibility, we adopt the cross-modal network following [35] and utilize the Swin Transformer [21]/BERT [20] for image/text modalities. Notably, our method can be used as a plug-and-play module, so we can transform any state-of-the-art cross-modal network and independent networks in our framework, and more analyses are provided in the experiments part. In the following subsections, we will first provide the exploration of modal imbalance, then introduce our framework and learning objectives.

A. Exploration of Modal Imbalance

We investigate the phenomenon of imbalance between image and text modalities, specifically focusing on differences in modality sufficiency. Proposition 2 in [16] indicates that modality sufficiency is correlated with the performance of the optimal model: greater insufficiency leads to poorer performance in single-modal models. Hence, we assess modality sufficiency based on single-modal retrieval performance. Specifically, we use the ResNet101 [36] and Swin Transformer (SwinT) [21] for training the image modality, and LSTM [37] and BERT [20] for the text modality. The results are shown in Table I, which shows that text consistently outperforms images in single-modal retrieval across all datasets, indicating that the sufficiency of the text modality is greater than that of the image modality. Consequently, we categorize text as the strong modality and image as the weak modality. Furthermore, Swin Transformer and BERT exhibit the best retrieval performance for images and text, respectively.

Additionally, to further study whether modals can sufficiently represent each other, we conducted experiments where

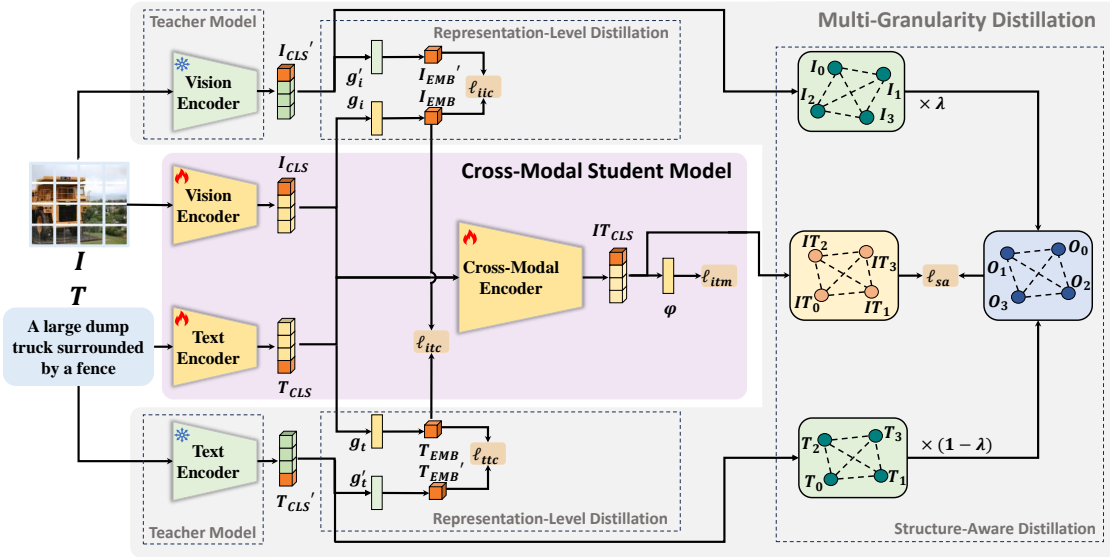


Fig. 3. Illustration of our framework. Expanding on the framework of cross-modal matching, we incorporate a single-modal teacher network. Our multi-granularity distillation includes representation-level distillation and structure-aware distillation, where the former optimizes the expressive capabilities of individual modalities via contrastive loss, and the latter enhances instance-level matching through structure-aware distillation.

the strong modality text distills the weak modality image (denoted as S2W@Image), and the weak modality distills the strong modality (denoted as W2T@Text). Specifically, in the S2W@Image setting, BERT serves as the teacher model and SWIN as the student model, with vision-language contrastive learning applied between them. The results are presented in Table II. We can find that: 1). The I2I performance of single-modal retrieval (SMR) is worse than that of S2W@Image, demonstrating that the strong modality, provides sufficient information to represent the weak modality; 2). The T2T performance of SMR is better than that of W2S@Text, demonstrating that the weak modality image cannot provide sufficient information to represent the strong modality text.

B. Cross-Modal Student Model

We use X-VLM [35] without the bounding box module as the cross-modal student model.

Vision and Language Encoders: Vision encoder produces fine-grained visual concept representations based on the SOTA Swin Transformer. As shown in Fig. 3, Swin Transformer splits an image $I \in \mathbb{R}^{C \times H \times W}$ into patches and flats to $I \in \mathbb{R}^{(C \times P^2) \times L_I}$, where $P \times P$ represents the patch resolution and $L_I = HW/P^2$ represents the number of patches following [38]. On the other hand, the language encoder maps the input text $T \in \mathbb{R}^{L_T \times d}$ to the same dimension subspace of the image with BERT, L_T is the words number, and d is the dimension of the common subspace. We recommend referring to [21, 38] for more detailed information.

Cross-Modal Encoder: It refers to the cross-attention transformer, i.e., the text-oriented cross-attention transformer, which aims to harness the efficacy of interaction layers to process visual and textual representations. As shown in Fig. 3, the cross-attention can be formulated as:

$$Att(T) = softmax\left(\frac{Q_T K_I^\top}{\sqrt{d_M}}\right) V_I, \quad (1)$$

where $Q_T = T W_{qT}$, $K_I = I W_{kI}$, and $V_I = I W_{vI}$ represent queries, keys, and values. $W_{qT} \in \mathbb{R}^{d_T \times d_M}$, $W_{kI} \in \mathbb{R}^{d_I \times d_M}$, and $W_{vI} \in \mathbb{R}^{d_I \times d_M}$ are learnable mapping matrixes. Besides, multi-head attention is composed of M parallel heads, and $d_M = d/M$. As a result, the other modal representations (i.e., query) are fused with the self-representations (i.e., keys and values) at each attention layer $Att(T)$ in Equation 1. Note that we directly adopt the pre-trained network [35] for fine-tuning.

Cross-Modal Matching: Cross-modal matching aims to learn consistent cross-modal representations, including vision-language contrastive learning (ITC) and vision-language matching (ITM). Following X-VLM [35], ITC preliminarily learns cross-modal consistency representations by treating all samples equally and performing contrastive learning across every pair of samples within the batch, thereby facilitating the initial filtering of similar sample pairs prior to vision-language fusion. In detail, we randomly sample a mini-batch of J pairs and calculate the in-batch image-to-text and text-to-image similarity. This process can be represented as:

$$\begin{aligned} l_{itc} &= \frac{1}{2} \mathbb{E}_{(I,T)} [CE(\mathbf{y}^{i2t}(I), \mathbf{p}^{i2t}(I)) + CE(\mathbf{y}^{t2i}(T), \mathbf{p}^{t2i}(T))], \\ p_k^{i2t}(I) &= \frac{\exp(d(I, T_k)/\tau)}{\sum_{j=1}^J \exp(d(I, T_j)/\tau)}, \\ p_k^{t2i}(T) &= \frac{\exp(d(T, I_k)/\tau)}{\sum_{j=1}^J \exp(d(T, I_j)/\tau)}. \end{aligned} \quad (2)$$

Here, $\mathbf{y}^{i2t}(I), \mathbf{y}^{t2i}(T) \in \{0, 1\}^J$ denote the matching ground-truth, where $y_j^{i2t} = 1$ if (I, T_j) is matched and $y_j^{i2t} = 0$ otherwise. $\mathbf{p}^{i2t}(I)$ and $\mathbf{p}^{t2i}(T)$ represent the in-batch image-to-text and text-to-image similarity, respectively. $d(I, T) = \cos(g_i(I_{CLS}), g_t(T_{CLS}))$ and $d(T, I) = \cos(g_t(T_{CLS}), g_i(I_{CLS}))$ denote the similarity, I_{CLS} and T_{CLS} are the [CLS] embeddings output by the visual and language encoders, respectively. g_t and g_i are linear transformations that map the [CLS] embeddings to a lower dimension. The instances in the same batch act as the negative anchor.

τ denotes the temperature scale parameter. CE is the cross-entropy loss.

Consistent with the X-VLM [35], ITM leverages hard negative sampling within the batch to predict whether a given image-text pair is a match, thereby enabling the learning of more refined margins. In detail, for each image in a mini-batch, we sample a hard negative text according to $p_k^{t2i}(T)$ in Equation 2. We use the [CLS] embedding output IT_{CLS} of the cross-modal encoder to predict the matching probability, and the loss can be represented as:

$$\ell_{itm} = \mathbb{E}_{(I,T)}[CE(\mathbf{y}^{match}, \mathbf{p}^{match})],$$

where \mathbf{y}^{match} is a 2-dimensional one-hot vector representing the ground-truth label, $\mathbf{p}^{match} = \phi(IT_{CLS})$ represents the matching prediction output by the binary classifier $\phi(\cdot)$. Therefore, the overall cross-modal loss can be formulated as: $\ell_{cr} = \ell_{itc} + \ell_{itm}$.

C. Multi-Granularity Distillation

We employ single-modal teacher models to guide the cross-modal model, implementing multi-granularity distillation that incorporates structural distillation [39] on top of representational distillation. This approach preserves the structural consistency of cross-modal representations within the latent space, functioning as a rebalancing mechanism to enhance cross-modal retrieval performance.

Single-Modal Teacher Models: We construct advantaged teacher models for two modalities, which aim to transfer structural knowledge for the cross-modal network in learning matching representations. Note that traditional single-modal contrastive learning usually pre-trains models pairwise, leading to some semantically similar instances being regarded as antagonistic pairs. Therefore, inspired by [40], we employ unsupervised prototype-aware contrastive learning for preserving the single-modal structure, which can distinguish intra- and inter-cluster pairs with a distance regularization. Given that labels may be available in practical scenarios, we investigate the impact of the teacher model under supervised learning. Furthermore, we examine the performance of pre-trained models, DINO [41] and T5 [42]. The results of both experiments are provided in Section IV of the supplementary material.

Representation-Level Distillation: At the single-modal representation level, we employ contrastive learning to narrow the expression gap between the student model and the teacher model, thereby mitigating the influence on single-modal representations during cross-modal matching. Specifically, with the randomly sampled batch pairs J , we calculate the similarity between the output of the visual model and the visual teacher model, as well as the output of the language model and the language teacher model. Each anchor instance's single-modal representations, paired with the corresponding teacher model output, form positive pairs, while the instances from

the batch form negative pairs. The optimization objectives are as follows:

$$\begin{aligned} \ell_{iic} &= \mathbb{E}_{(I,I)}[CE(\mathbf{y}^{i2i}(I), \mathbf{p}^{i2i}(I))], \\ \ell_{ttc} &= \mathbb{E}_{(T,T)}[CE(\mathbf{y}^{t2t}(T), \mathbf{p}^{t2t}(T))], \\ p_k^{i2i}(I) &= \frac{\exp(d(I, I_k)/\tau)}{\sum_{j=1}^J \exp(d(I, I_j)/\tau)}, \\ p_k^{t2t}(T) &= \frac{\exp(d(T, T_k)/\tau)}{\sum_{j=1}^J \exp(d(T, T_j)/\tau)}, \end{aligned} \quad (3)$$

where $\mathbf{y}^{i2i}(I), \mathbf{y}^{t2t}(T) \in \{0, 1\}^J$ denote the matching ground-truth, which are defined similarly to $\mathbf{y}^{i2t}(I), \mathbf{y}^{t2i}(T)$. $\mathbf{p}^{i2i}(I)$ denotes the in-batch image-to-image similarity between the outputs of student and teacher model. And $\mathbf{p}^{t2t}(T)$ is defined similarly. The ℓ_{iic} and ℓ_{ttc} aims to align the single-modal representations of each instance with their corresponding teacher model outputs. This ensures that the learning process does not disrupt the internal similarity relationships within the single-modal representations, thereby facilitating the effective transfer of structural knowledge to the cross-modal model.

Structure-Aware Distillation: As shown in Fig. 3, we calculate the relational similarity matrices S_I, S_T, S_{IT} for two modal independent networks and the cross-modal network, by computing the distances of data examples in the mini-batch. Using the S_I as an example: $S_I(i, j) = \cos(I_{iCLS}, I_{jCLS})$. Consequently, we can acquire the optimal teacher matrix by fusing two teachers' matrices $S_O = \lambda S_I + (1 - \lambda) S_T$, where λ is a learnable parameter aimed at finding the optimal modality fusion. For measuring the teacher-student relational consistency, we adopt the Mean Absolute Error (MAE) between two matrices, which aims to find subtler differences by considering matrices as high-dimensional objectives rather than simple aggregations. Consequently, given S_O, S_{IT} , the structure-aware distillation is defined as:

$$\ell_{sa} = \frac{1}{J} \sum_{m=1}^J \sum_{n=1, n \neq m}^J |S_O(m, n) - S_{IT}(m, n)|.$$

This effectively alleviates the disruption of structural information during modal alignment. Therefore, the overall multi-granularity distillation can be represented as: $\ell_{md} = \ell_{iic} + \ell_{ttc} + \ell_{sa}$.

In summary, we combine cross-modal matching loss and distillation objectives to study the common cross-modal representations, which can reduce the cross-modal gap and preserve structure simultaneously. The overall loss can be formulated as: $L = \ell_{cr} + \ell_{md}$.

IV. EXPERIMENTS

A. Data Description

To verify the effectiveness of our method, we conduct experiments on three datasets. In detail, **MS-COCO** [43] contains 123,287 images, including 82,783 training images and 40,504 validation images, each labeled with 5 captions. Following [44], we use the splits of 5,000 images for validation, 1,000/5,000 images for testing, and the rest for training. **FLICKR30K** [45] consists of 31,000 images, and each image is associated with 5 captions. The dataset is split into 29,000 training images, 1,000 validation images, and 1,000 testing

TABLE III
CROSS-MODAL RETRIEVAL PERFORMANCE COMPARISON. EVALUATION CRITERIA IS R@A. THE METHOD WITH “+” SIGN IS OUR METHOD.

Methods	MS-COCO (1K)						MS-COCO (5K)						FLICKR30K						Vizwiz					
	I2T			T2I			I2T			T2I			I2T			T2I			I2T			T2I		
	@1	@5	@10	@1	@5	@10	@1	@5	@10	@1	@5	@10	@1	@5	@10	@1	@5	@10	@1	@5	@10	@1	@5	@10
VSE++	49.5	81.0	90.0	38.1	73.3	85.1	39.0	67.9	79.5	29.3	59.1	72.4	31.6	59.3	71.7	21.6	50.7	63.8	35.1	58.1	65.4	25.3	48.1	58.4
SCAN	72.7	94.8	98.4	58.8	88.4	94.8	48.3	82.0	89.1	38.1	68.9	80.2	60.0	83.9	90.7	37.7	66.3	76.0	34.9	60.9	72.8	24.8	48.6	59.0
IMRAM	76.7	95.6	98.5	61.7	89.1	95.0	52.8	82.6	89.8	39.1	68.6	77.9	74.1	93.0	96.6	53.9	79.4	87.2	42.5	67.5	78.6	27.6	52.5	63.6
SGRAF	79.6	96.2	98.5	63.2	90.7	96.1	57.6	83.5	91.5	41.9	71.3	81.2	77.8	94.1	97.4	58.5	83.0	88.8	43.9	73.4	80.1	28.8	54.4	64.2
GSMN	78.4	96.4	98.6	63.3	90.1	95.7	55.2	81.3	86.2	37.2	68.3	77.3	76.4	94.3	97.3	57.4	82.3	89.0	43.3	72.4	79.3	26.9	53.2	63.8
VSRN	69.5	92.3	97.0	54.1	83.4	90.2	53.0	81.1	89.4	40.5	70.6	81.1	58.0	86.1	91.6	46.9	77.0	85.1	34.8	63.1	73.5	26.3	52.5	64.1
NAAF	78.1	96.1	98.6	63.5	89.6	95.3	58.9	85.2	92.0	42.5	70.9	81.4	79.6	96.3	98.3	59.3	83.9	90.2	44.1	69.9	78.0	31.0	54.8	64.2
ALBEF*	80.1	96.9	99.0	68.3	92.5	97.1	59.7	85.8	92.3	46.1	75.8	84.9	63.2	87.4	93.5	48.5	73.1	80.7	47.7	70.4	79.6	34.3	56.3	65.7
BLIP*	81.4	96.7	99.2	67.7	92.1	96.3	57.8	84.1	91.2	43.4	72.8	82.7	65.0	89.5	94.1	52.2	81.9	89.2	46.2	73.2	82.2	37.2	64.0	74.5
CYCLIP*	78.4	96.0	98.9	65.7	91.5	96.6	53.0	81.0	89.3	41.0	71.7	82.7	77.1	94.1	97.8	61.4	88.1	92.2	52.4	79.7	87.2	39.2	70.5	79.6
UMT	81.5	97.3	99.1	69.2	93.2	97.2	58.4	85.0	92.7	43.6	74.5	84.3	77.6	94.8	97.9	61.9	89.8	93.3	52.7	81.2	87.6	41.8	71.7	80.4
X-VLM*	81.1	97.7	99.5	70.3	94.7	97.5	55.2	86.6	93.3	44.8	76.9	86.0	78.1	96.0	98.8	65.4	90.4	94.9	57.7	82.8	89.3	44.2	73.5	81.7
X-VLM*+	87.3	98.7	99.8	73.4	94.8	97.7	66.6	90.5	95.5	49.9	79.5	87.6	85.6	98.1	99.3	73.3	93.0	96.1	63.9	85.7	90.6	50.7	75.5	83.7

TABLE IV
SINGLE-MODAL RETRIEVAL PERFORMANCE COMPARISON. EVALUATION CRITERIA IS NDCG@A. THE METHOD WITH “+” SIGN IS OUR METHOD.

Methods	MS-COCO (1K)						MS-COCO (5K)						FLICKR30K						Vizwiz					
	I2I			T2T			I2I			T2T			I2I			T2T			I2I			T2T		
	@10	@20	@50	@10	@20	@50	@10	@20	@50	@10	@20	@50	@10	@20	@50	@10	@20	@50	@10	@20	@50	@10	@20	@50
VSE++	62.5	64.2	66.5	67.9	69.4	70.2	43.2	45.2	48.5	43.8	45.5	48.6	59.0	60.9	64.0	63.3	64.0	64.2	54.2	55.6	56.7	57.8	58.6	59.3
SCAN	64.1	65.1	66.7	69.9	70.3	69.5	59.7	61.2	63.4	63.5	64.8	65.8	52.6	55.0	59.1	38.7	42.6	47.6	55.4	57.4	61.1	52.0	54.0	56.5
IMRAM	65.1	66.6	68.6	70.2	70.5	70.1	60.4	62.0	64.5	64.4	66.1	67.9	61.1	62.8	65.7	66.8	66.4	65.5	59.5	61.5	65.0	61.2	62.4	63.1
SGRAF	62.5	64.0	66.3	61.3	63.4	65.0	56.7	58.3	60.8	53.8	56.1	58.4	58.7	60.8	64.3	57.7	60.2	62.9	59.1	61.1	64.7	56.3	59.3	62.9
GSMN	61.6	63.3	65.9	59.7	61.1	61.5	55.1	56.7	59.1	49.1	50.7	53.1	60.1	61.9	65.2	53.6	56.4	59.7	56.4	57.3	58.6	52.5	56.3	55.9
VSRN	52.9	55.1	58.8	58.3	57.6	56.9	45.0	46.3	49.1	60.3	58.4	56.3	61.7	63.4	66.4	67.8	67.4	66.3	62.0	63.7	66.7	59.8	60.8	61.6
NAAF	63.7	64.8	66.8	70.7	70.9	72.1	59.5	60.8	62.8	65.3	66.6	68.6	60.5	62.2	65.2	71.1	69.9	68.2	59.1	61.2	64.7	61.7	61.8	62.9
ALBEF*	68.2	69.1	70.9	49.6	50.6	51.5	64.3	66.1	68.3	47.5	49.7	51.1	60.8	62.3	65.2	70.0	68.4	66.2	61.8	63.3	66.2	63.7	63.0	62.1
BLIP*	68.1	68.9	70.1	64.4	63.5	62.4	64.0	65.5	67.6	59.0	59.3	59.5	62.6	64.2	67.0	65.7	64.4	62.9	63.0	64.9	68.0	58.3	58.8	59.6
CYCLIP*	67.3	68.2	69.0	68.9	68.4	67.3	62.7	64.0	66.3	63.4	64.5	66.8	61.9	63.3	66.2	63.4	64.0	63.9	63.0	64.5	67.4	56.0	58.2	59.9
UMT	67.5	68.8	70.7	68.1	65.9	64.4	65.2	68.0	67.7	57.3	57.2	66.3	62.7	64.2	66.9	66.8	65.3	63.9	64.0	65.5	67.4	59.0	60.2	61.9
X-VLM*	69.2	70.0	71.1	72.8	72.9	72.3	65.3	66.8	68.8	65.5	66.7	67.8	64.1	65.2	67.7	72.3	70.9	68.7	64.4	66.1	68.9	64.8	64.9	65.0
X-VLM*+	70.1	70.9	72.1	73.5	73.4	72.5	66.3	67.8	69.8	65.8	67.0	68.0	64.3	65.4	67.8	73.0	71.4	68.9	64.4	66.1	68.9	67.2	66.5	65.5
SWIN	66.5	67.7	69.1	-	-	-	61.5	63.2	65.7	-	-	-	64.1	64.5	67.9	-	-	-	50.3	53.9	57.3	-	-	-
BERT	-	-	-	71.1	71.1	70.2	-	-	-	63.4	64.8	66.1	-	-	-	71.6	70.2	68.4	-	-	-	55.5	57.5	59.6

images following [44]. **Vizwiz** [46] consists of 39,181 images originating from people who are blind that are each paired with 5 captions. The dataset is split into 23,431 training instances, 7,750 validation instances, and 8,000 testing instances.

B. Baselines and Evaluation Protocol

For comparison methods, we evaluate four types of state-of-the-art approaches: 1). Dual-encoder retrieval methods, including VSE++ [8], SCAN [6], IMRAM [47], SGRAF [9], GSMN [48], VSRN [49], and NAAF [50]. 2). Transformer-based retrieval methods, including ALBEF [31], BLIP [14], and X-VLM [35]. 3). Single-modal models, e.g., Swin Transformer [21] and BERT [20]. 4). Imbalance multi-modal learning methods, i.e., CYCLIP [51] and UMT [52]. Note that CYCLIP incorporates regularizers about single-modal and cross-modal structural information into the cross-modal contrastive item in the form of a multi-task loss, and UMT optimizes the cross-modal model by distilling the instance’ relations learned by the single-modal model. Since the baselines utilize different backbones, we conducted an experiment that integrates our method with these baselines in a plug-and-play manner to ensure a fair comparison. The detailed results are provided in Section III of the supplementary.

In our experiments, we focus on three tasks: 1). cross-modal retrieval, 2). single-modal retrieval, and 3). mixed retrieval. Detailed implementation of these three retrieval tasks is provided in Section I of the supplementary. We evaluate performance using the recall at A (R@A) metric, which is commonly adopted in most cross-modal retrieval methods for both image-to-text (I2T) and text-to-image (T2I) retrieval tasks [8, 27, 31, 35]. During cross-modal retrieval, the corresponding captions of the given image or corresponding images of the given caption are expected, while in single-modal retrieval, what we need are relatively similar instances [53, 54, 55]. As a result, the R@A metric as a binary correlation that only examines whether the retrieval results are relevant is not suitable for single-modal and mixed retrieval (I2IT and T2IT). Following [27, 56], a more comprehensive metric NDCG@A [56] is adopted instead to promote the items with higher relevance scores to appear in better ranking positions. Note that the ranking is computed using text similarity scores (i.e., ROUGE-L [57]) between a sentence and the sentences associated with a certain image.

C. Implementation Details

Our multi-granularity distillation approach can be seen as a plug-and-play module, and we can incorporate other cross-

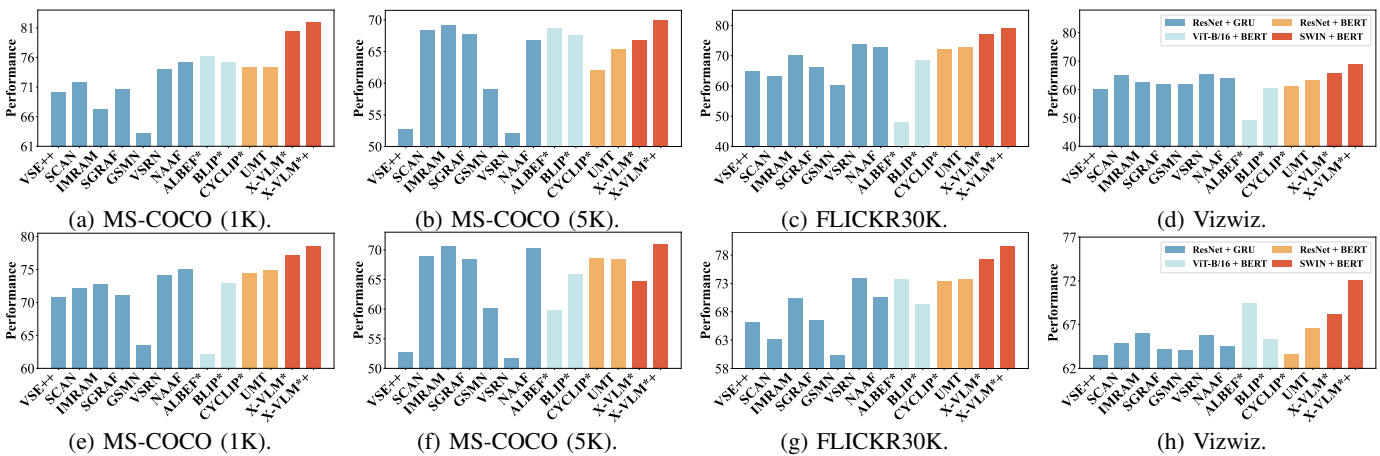


Fig. 4. The results of I2IT@10 (a–d), T2IT@10 (e–h) of mixed retrieval task. The method with “+” sign, i.e., X-VLM*+, is our method.

modal retrieval models as student models to validate the generalization capability of the multi-granularity distillation. Without any loss of generality, we choose X-VLM without the bounding box module as the student model to validate the effectiveness of our method, and in subsequent experiments, we will extend this module to traditional cross-modal retrieval and visual-language pre-training models, to further demonstrate its generalization. X-VLM consists of a vision encoder which is initialized using Swin Transformer/12 [38], a language encoder which is initialized using the first 6 layers of the BERTbase [20], and text-oriented cross-attention transformer are initialized using the last 6 layers of the BERTbase. In total, our model has 214.4M parameters for training. We also study the impact of using larger models such as CLIP for training, as discussed in Section V of the supplementary. The images’ resolution is 384×384 as input. For text input, the maximum number of tokens is 30. We use the AdamW [58] optimizer with a weight decay of 0.1. The learning rate is warmed up to 0 from $3e-5$ in the first epoch and decayed to $1e-5$ following a cosine schedule. The momentum parameter for updating the momentum model is 0.995. The model is trained for 10 epochs with a total batch size of 36 on 6 NVIDIA A6000 GPUs.

D. Retrieval Results

Cross-Modal Retrieval: I2T and T2I retrievals are considered to evaluate cross-modal retrieval performance. Table III records the results on four public datasets. To ensure experimental fairness, we exclusively utilized the given datasets for model training, rather than pre-training additional data as the large-scale models (i.e., ALBEF, X-VLM, CYCLIP, and BLIP) do. So we retrained the ALBEF and other large models from scratch (marked with “*”), thus causing different results compared to the original paper. Results indicate: 1). Our X-VLM*+ outperforms the best cross-modal retrieval method, i.e., X-VLM, on I2T Recall@1 and T2I Recall@1 by 6.2/11.4/7.5/6.2 and 3.1/5.1/7.9/6.5 respectively, on four datasets. This phenomenon reveals that structure preservation can also enhance the learning of cross-modal consistent representations by bringing image (i.e., weak modality) representations closer to the text (i.e., strong modality) ones without

TABLE V
ABLATION STUDIES PERFORMANCE COMPARISON.
EVALUATION CRITERIA ARE R@A AND NDCG@A.

Methods	FLICKR30K						Vizwiz					
	I2T			T2I			I2T			T2I		
	@1	@5	@10	@1	@5	@10	@1	@5	@10	@1	@5	@10
w/o <i>litc</i>	84.8	97.6	99.3	71.4	92.6	95.7	61.4	85.4	90.3	47.7	75.1	83.1
w/o <i>litm</i>	78.6	97.0	98.9	65.7	90.3	94.5	60.2	85.3	90.8	46.3	74.7	83.1
w/o <i>liic</i>	83.5	96.9	98.8	68.0	91.6	95.3	59.4	83.2	89.8	45.1	73.6	81.9
w/o <i>lttc</i>	84.1	97.1	99.1	67.7	91.6	94.8	61.1	85.0	90.4	46.9	75.2	83.4
w/o <i>lsa</i>	81.2	95.8	99.0	67.3	90.9	95.0	58.1	82.8	89.9	45.2	73.9	82.3
Ours	85.6	98.1	99.3	73.3	93.0	96.1	63.9	85.7	90.6	50.7	75.5	83.7
	I2I			T2T			I2I			T2T		
	@10	@20	@50	@10	@20	@50	@10	@20	@50	@10	@20	@50
w/o <i>litc</i>	63.1	64.5	67.2	72.0	70.9	68.4	63.1	64.8	67.5	65.3	64.9	64.6
w/o <i>litm</i>	64.3	65.6	68.2	73.7	72.1	69.9	64.0	65.7	68.5	66.0	65.8	65.4
w/o <i>liic</i>	63.9	65.0	67.2	72.9	70.9	68.1	63.8	65.8	68.4	64.4	64.8	65.2
w/o <i>lttc</i>	64.1	65.0	67.0	72.7	70.3	67.2	64.2	65.9	68.2	65.3	65.4	65.3
w/o <i>lsa</i>	63.5	64.7	66.9	72.7	70.0	67.3	64.0	65.5	68.1	64.6	65.1	65.4
Ours	64.3	65.4	67.8	73.0	71.4	68.9	64.4	66.1	68.9	67.2	66.5	65.5
	I2IT			T2IT			I2IT			T2IT		
	@10	@20	@50	@10	@20	@50	@10	@20	@50	@10	@20	@50
w/o <i>litc</i>	78.7	73.9	68.3	77.2	75.3	72.3	66.8	65.2	62.3	68.5	67.8	66.2
w/o <i>litm</i>	82.7	79.0	74.4	82.9	79.2	74.5	75.1	74.8	71.0	78.3	75.0	71.2
w/o <i>liic</i>	78.0	73.7	68.4	77.3	75.3	72.0	66.8	65.6	63.0	68.4	68.0	67.0
w/o <i>lttc</i>	78.2	73.5	68.2	76.4	75.7	72.2	67.0	66.0	62.5	67.7	67.6	66.5
w/o <i>lsa</i>	77.7	73.3	67.3	77.1	74.8	71.9	66.3	65.1	62.8	68.5	68.1	67.1
Ours	79.1	74.0	68.8	79.6	76.7	72.5	68.8	66.3	63.1	72.0	70.2	67.6

compromising the original single-modal structure. 2). Our X-VLM*+ outperforms CYCLIP* (multi-task optimization) in all settings, which emphasizes that multi-granularity distillation emerges as a relatively superior strategy. 3). The performance of large-scale pre-trained models is limited when data is limited, e.g., ALBEF* and X-VLM* exhibit competitive performance with dual-encoder models like SGRAF and NAAF. **Single-Modal Retrieval:** Conducting two tasks, i.e., I2I and T2T, we aim to verify whether traditional vision-language models can preserve structural information of single modality after cross-modal representation learning. Table IV exhibits the single-modal retrieval performance between traditional models and vision-language pre-training models. Note that “SWIN” and “BERT” represent directly utilizing unsupervised prototype-aware contrastive learning for training, and then test the single-modal retrieval performance. Experiment results

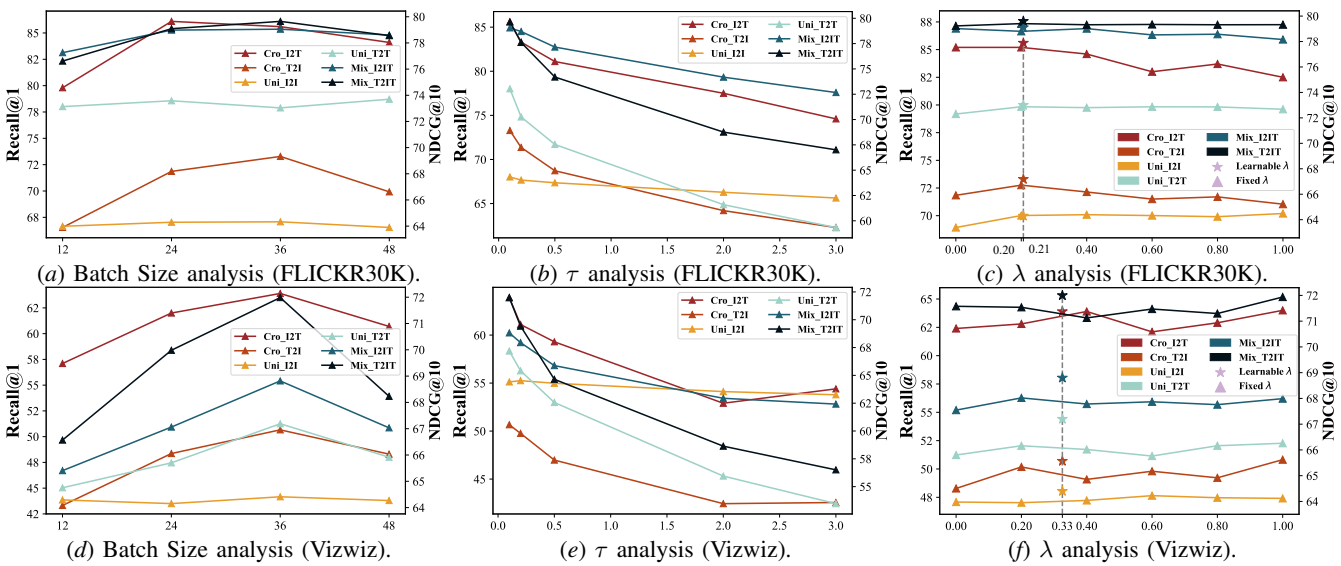


Fig. 5. **Parameter analyses.** We verify the influence of parameters our method under Flickr30K and Vizwiz datasets.

TABLE VI
COMPARE MAE WITH MSE, KL, AND WD. EVALUATION CRITERIA ARE R@A AND NDCG@A.

Methods	Flickr30K						Vizwiz					
	I2T			T2I			I2T			T2I		
	@1	@5	@10	@1	@5	@10	@1	@5	@10	@1	@5	@10
X-VLM*+ (MAE)	85.6	98.1	99.3	73.3	93.0	96.1	63.9	85.7	90.6	50.7	75.5	83.7
X-VLM*+ (MSE)	84.0	97.6	99.0	69.6	91.7	95.1	60.9	84.4	90.6	<u>48.5</u>	75.7	<u>83.4</u>
X-VLM*+ (KL)	7.3	22.8	36.3	1.8	7.1	11.7	8.9	25.3	<u>38.5</u>	4.1	14.2	23.4
X-VLM*+ (WD)	85.8	97.5	99.0	<u>70.7</u>	<u>92.1</u>	<u>95.2</u>	<u>63.1</u>	<u>85.1</u>	90.6	48.4	75.4	83.7
	I2I			T2T			I2I			T2T		
	@10	@20	@50	@10	@20	@50	@10	@20	@50	@10	@20	@50
X-VLM*+ (MAE)	64.3	65.4	<u>67.8</u>	73.0	71.4	68.9	64.4	66.1	68.9	67.2	66.5	65.5
X-VLM*+ (MSE)	63.8	<u>65.2</u>	68.0	73.0	<u>70.9</u>	68.1	64.1	<u>65.9</u>	<u>68.8</u>	<u>65.8</u>	<u>65.8</u>	<u>65.6</u>
X-VLM*+ (KL)	61.5	63.1	66.1	71.5	70.0	68.0	63.2	64.9	68.1	61.6	62.5	63.2
X-VLM*+ (WD)	<u>64.0</u>	65.1	<u>67.8</u>	<u>72.8</u>	<u>70.8</u>	<u>68.2</u>	<u>64.3</u>	<u>65.9</u>	<u>68.8</u>	<u>65.8</u>	<u>65.8</u>	65.7
	I2IT			T2IT			I2IT			T2IT		
	@10	@20	@50	@10	@20	@50	@10	@20	@50	@10	@20	@50
X-VLM*+ (MAE)	79.1	74.0	68.8	79.6	76.7	<u>72.5</u>	68.8	66.3	<u>63.1</u>	72.0	70.2	67.6
X-VLM*+ (MSE)	<u>78.2</u>	73.2	68.3	78.4	<u>76.4</u>	72.1	66.9	66.1	63.7	67.5	67.5	66.6
X-VLM*+ (KL)	48.4	48.8	50.3	75.1	73.5	71.4	51.3	51.8	53.5	67.6	66.7	65.5
X-VLM*+ (WD)	78.0	74.5	<u>68.6</u>	<u>78.5</u>	75.6	73.2	<u>67.1</u>	<u>66.2</u>	63.7	<u>68.0</u>	<u>67.9</u>	<u>66.8</u>

indicate that: 1). Our X-VLM*+ performs better than nearly all cross-modal retrieval comparison methods in both I2I and T2T retrievals on four datasets. It even outperforms the best single-modal retrieval methods, on I2I NDCG@10 and T2T NDCG@10 by 3.6/4.8/0.2/14.1 and 2.4/2.4/1.4/11.7 respectively. 2). Although several cross-modal retrieval methods can improve the I2I retrieval, e.g., ALBEF* increases the NDCG@10 compared with Swin Transformer (1.7/2.8 on MS-COCO (1K)/MS-COCO (5K)), almost all methods exhibit varying degrees of performance decline in T2T retrieval, with the improvement in the weak modality's performance being less pronounced than the decline in the strong modality's performance. For instance, ALBEF* decreases 21.5/15.9 of T2T NDCG@10 on MS-COCO (1K) and MS-COCO (5K) datasets compared with BERT, with only 1.7/2.8 promotion of I2I NDCG@10 over Swin Transformer. 3). Our X-VLM*+ performs better than CYCLIP*, which reveals that adaptively distillation is superior to learning structure-preserving repre-

sentations by better incorporating structure consistency.

Mixed Retrieval: It further simulates the real task that retrieves all modal instances with single-modal query (i.e., image or text). Figures 4 (a–d) and (e–h) represent the results of I2IT NDCG@10 and T2IT NDCG@10, respectively, which reveal that: 1). As the position of the expected instances A increases, the performance of several models degrades. Since the retrieval library has limited capacity, even though most expected instances rank high, the remaining results may have low similarity, causing a decline in the similarity-based metric NDCG. 2). Our X-VLM*+ outperforms nearly all comparison methods on four datasets, which validates that our method can effectively learn cross-modal consistent and structure-preserving representations simultaneously. More comprehensive experimental results can be found in Section II of the supplementary due to space limitation.

TABLE VII
PERFORMANCE OF CROSS-MODAL, SINGLE-MODAL, AND MIX-MODAL RETRIEVAL WITH PRE-TRAINED MODELS ON FLICKR30K DATASET. EVALUATION CRITERIA ARE R@A AND NDCG@A. THE METHOD WITH “+” SIGN IS OUR METHOD.

Methods	Cross-Modal Retrieval						Single-Modal Retrieval						Mixed Retrieval					
	I2T			T2I			I2I			T2T			I2IT			T2IT		
	@1	@5	@10	@1	@5	@10	@10	@20	@50	@10	@20	@50	@10	@20	@50	@10	@20	@50
X-VLM	95.5	99.8	100.0	85.3	97.0	98.5	64.1	65.6	68.3	74.9	72.8	70.3	85.5	79.2	71.4	82.0	78.6	74.1
X-VLM+	96.6	100.0	100.0	86.8	97.7	99.0	64.3	65.7	68.5	75.3	73.4	70.9	86.5	79.6	71.6	82.4	79.2	74.7









Query	Result	Ours	VSRN
 (a)		1: A mother decides to take her child on a piggyback ride outside their apartment complex. ✓ 2: A woman gives a small child a piggyback ride. ✓ 3: A young woman is giving a baby a ride on her shoulders. ✓ 4: A baby boy in a blue and white striped shirt is sitting on his mothers shoulders. ✓ 5: A female child is on the shoulders of parent. ✗	1: Mother tending to her child. ✗ 2: A woman gives a small child a piggyback ride. ✓ 3: A woman with short hair holds a small baby in her arms. ✗ 4: A little girl is holding a little boy on her lap. ✗ 5: A young blond girl holds a smaller blond boy in her lap outdoors. ✗
 (b)		1: Three men and two women stand facing the ocean from the shore on a sunny day. ✓ 2: Two ladies and three men looking at the ocean. ✓ 3: Five people look out toward the ocean. ✓ 4: A group of people stand in the sand looking out at the water. ✓ 5: Five people standing in front of a body of water. ✓	1: A group of people stand in the sand looking out at the water. ✓ 2: A group of people playing on a beach. ✗ 3: Three men and two women stand facing the ocean from the shore on a sunny day. ✓ 4: A family fishes off the edge of a beach. ✗ 5: Five people look out toward the ocean. ✓
(c) A black and white dog is running in a grassy garden surrounded by a white fence.			
(d) Five people standing in front of a body of water.			

Fig. 6. Qualitative results of cross-modal retrieval. For each image query, we show the top-5 ranked sentences. For each sentence query, we show the top-3 ranked images, ranking from left to right (Best viewed in green). The examples are sampled from the FLICKR30K dataset.

E. Ablation Study

Furthermore, we conduct ablation studies to validate the effectiveness of each module: cross-modal matching (l_{itc} and l_{itm}), representation-level distillation losses (l_{iic} and l_{ttc}) and the structure-aware distillation module (l_{sa}). Due to the large size of the MS-COCO dataset, we focus on the smaller FLICKR30K and Vizwiz datasets in subsequent experiments. Analyzing the results in Table V reveals the following observations: 1). The removal of l_{itm} leads to a significant decline in cross-modal retrieval performance, although single-modal retrieval performance improves, further confirming the negative impact of cross-modal consistency learning on single-modal retrieval. 2). Compared to the removal of l_{itc} and l_{itm} , the elimination of l_{sa} results in the poorest retrieval performance, indicating that preserving instance structure information more effectively promotes representation learning both across modalities and within each modality.

F. Sensitivity to Parameters

To verify the influence of parameters, we conduct more experiments by tuning several important parameters: 1). batch size J ; 2). temperature parameter τ ; 3). hyper-parameter λ .

Influence of Batch Size: We incorporate batch with different sizes, i.e., $\{12, 24, 36, 48\}$ into the proposed model, to empirically investigate the impact of batch size on performance. The performance in Fig. 5 (a) and (d) first increases and then decreases, indicating that a larger batch size can consider more neighbor information, but an oversized one may introduce noisy information.

Influence of Temperature Parameter τ : To explore the influence of the temperature scale parameter, we tune the $\tau \in \{0.1, 0.2, 0.5, 2, 3\}$ to conduct more experiments. Fig. 5 (b) and (e) record the results. We find that the retrieval results are the best when $\tau = 0.1$ on two datasets. This indicates that the points have few similar neighbors, which can promote the learning of structure-aware representations.

Influence of Hyper-Parameter λ : To investigate the importance of vision and language modalities in Equations 2 and 3, Figures 5 (c) and (f) present the results under fixed $\lambda \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ and learnable λ . We have observed the following phenomenons: 1). All five-star markers are positioned above the triangles on the same-colored lines, which indicates that the learnable λ effectively contributes to the discovery of the most optimal modality fusion for the model. 2). Specifically, on both the FLICKR30K and Vizwiz datasets, the ultimate values of learnable λ are 0.21 and 0.33,










Result Query	Ours	SCAN
 (a)	1: A group of volleyball girls high fiving each other in a gym. (0.9999) 2: Females who are on team usa jump in the air excitedly. (0.9527) 3: Two female members of team usa performing a jump high five surrounded by two other female members. (0.8602) 4: Two volleyball girls team both jumping for a double high five. (0.8039) 5: 0.6466 	1: Two teammates from the usa team are jumping in the air to give high fives while two teammates watch. (0.9999) 2: A female soccer team wearing green black and white uniforms is stretching before a game. (0.8530) 3: A team plays ball in the sand as a few spectators look on from the stands. (0.7406) 4: 0.6051  5: A group of male and female cheerleaders create three formations. (0.5827)
 (b)	1: A boy with a winged bug perched on his nose. (0.9999) 2: Boy that is outside with a green bug sitting on his nose. (0.9328) 3: A boy poses with a large green insect on his nose. (0.8275) 4: Young boy smiles as an extremely large kelly green fly perches on his nose. (0.7841) 5: 0.6710 	1: Young boy smiles as an extremely large kelly green fly perches on his nose. (0.9999) 2: A young smiling child holds his toy alligator up to the camera. (0.8584) 3: A man in a green hard hat and yellow safety vest frowns. (0.8081) 4: 0.6692  5: A woman is wearing a green red and white shirt. (0.7547)
(c) A girl in a polka dotted blue jean dress walks barefoot on a balance beam.	1: 0.9814  2: A girl in a jean dress is walking along a raised balance beam. (0.9437) 3: A little girl is walking on a balance beam. (0.8247) 4: The little girl is walking a beam in gymnastics class. (0.7963) 5: A little girl balances on a gymnastics beam. (0.7746)	1: A boy in a blue t shirt is airborne on a skateboard in a skate park. (0.8789) 2: A little girl is walking on a balance beam. (0.8697) 3: A man with a mask on sitting at a table with a plate in front. (0.8593) 4: 0.8616  5: A girl in a jean dress is walking along a raised balance beam. (0.8419)
(d) A female harp player peers through the middle of her instrument while performing.	1: 0.9999  2: A woman playing a harp. (0.8695) 3: A smiling woman playing the harp. (0.8400) 4: A lady with dark hair is playing a harp. (0.8378) 5: A pretty woman plays a harpsichord. (0.7657)	1: A woman is sitting with a basket of cloth surrounded by cloth. (0.8364) 2: A woman with long curly hair is standing at a podium and speaking. (0.8348) 3: A lady with blond hair and glasses works hard to clean up her table. (0.8327) 4: A woman acts out a dramatic scene in public behind yellow caution tape. (0.8299) 5: A woman working on her computer in front of a bright yellow wall. (0.8236)

Fig. 7. **Qualitative results of mixed retrieval.** For each query, we show the top-5 ranked instances, including images and sentences (Correct results viewed in green). The examples are sampled from the FLICKR30K dataset.

respectively. This indicates that the language modality (i.e., the strong modality) plays a more pivotal role in information provision during the distillation process.

G. Influence with Different Distillation Measurements

To explore the effectiveness of graph-matching criteria, we conducted more experiments. To be specific, we replace the MAE distance with other distance measurements, i.e., the Mean Squared Error (MSE), Kullback-Leibler divergence (KL), and Wasserstein distance (WD). The results in Table VI reveal that our graph-matching criteria, independent of specific distance metrics, demonstrate excellent flexibility, as observed in the MAE, MSE, and WD outcomes. In contrast, the KL method is not conducive to cross-modal retrieval.

H. Generalization of Multi-Granularity Distillation

To assess the generalization ability of our multi-granularity distillation strategy, we extended experiments to include the retrieval methods SCAN and VSRN based on dual encoders, along with the transformer-based retrieval method ALBEF. These experiments were conducted on the FLICKR30K and Vizwiz datasets. SCAN adopts a local representation matching approach, while VSRN introduces a global semantic reasoning module alongside attention to region relationships. As dual encoder methods do not learn modal fusion representations, we introduced the structure-aware distillation module for each single modality, aiming to enhance instance-level modality matching by leveraging the representation structure information from the single-modal teacher model. Results in Table VIII show a significant performance improvement when

TABLE VIII
PERFORMANCE OF THE MULTI-GRANULARITY DISTILLATION STRATEGY ON DIFFERENT RETRIEVAL MODELS (MARKED WITH "+"). EVALUATION CRITERIA ARE R@A AND NDCG@A.

Methods	FLICKR30K						Vizwiz					
	I2T			T2I			I2T			T2I		
	@1	@5	@10	@1	@5	@10	@1	@5	@10	@1	@5	@10
SCAN*	60.0	83.9	90.7	37.7	66.3	76.0	34.9	60.9	72.8	24.8	48.6	59.0
SCAN*+	61.0	86.5	92.5	40.8	72.3	81.2	38.3	64.4	73.4	27.6	52.9	62.8
VSRN*	58.0	86.1	91.6	46.9	77.0	85.1	34.8	63.1	73.5	26.3	52.5	64.1
VSRN*+	62.1	86.3	92.1	47.4	77.2	85.3	35.9	64.3	73.7	28.1	55.1	67.4
ALBEF*	63.2	87.4	93.5	48.5	73.1	80.7	47.7	70.4	79.6	34.3	56.3	65.7
ALBEF*+	65.8	88.6	94.8	50.7	75.3	82.6	50.2	73.3	82.0	36.9	61.2	70.1
	I2I			T2T			I2I			T2T		
	@10	@20	@50	@10	@20	@50	@10	@20	@50	@10	@20	@50
SCAN*	52.6	55.0	59.1	38.7	42.6	47.6	55.4	57.4	61.1	52.0	54.0	56.5
SCAN*+	53.4	55.6	59.6	45.6	49.5	54.3	58.5	60.5	64.0	53.4	56.8	60.4
VSRN*	61.7	63.4	66.4	67.8	67.4	66.3	62.0	63.7	66.7	59.8	60.8	61.6
VSRN*+	63.5	64.9	67.3	68.0	67.8	66.8	62.6	64.4	67.5	60.2	61.4	62.4
ALBEF*	60.8	62.3	65.2	70.0	68.4	66.2	61.8	63.3	66.2	63.7	60.2	62.1
ALBEF*+	60.8	62.4	65.2	72.2	70.3	68.1	61.8	63.4	66.4	65.3	64.5	63.4
	I2IT			T2IT			I2IT			T2IT		
	@10	@20	@50	@10	@20	@50	@10	@20	@50	@10	@20	@50
SCAN*	63.2	60.8	59.2	63.2	60.8	59.3	64.9	62.3	60.0	64.9	62.4	60.2
SCAN*+	67.1	65.5	64.4	67.1	65.6	64.5	65.8	64.7	62.4	71.4	70.0	65.5
VSRN*	73.8	71.6	68.5	73.9	71.7	68.7	65.4	63.8	63.4	65.7	64.2	63.7
VSRN*+	77.4	73.4	69.3	79.5	76.5	72.5	69.0	65.9	65.1	72.0	70.1	66.6
ALBEF*	48.1	48.3	49.7	73.8	71.1	67.9	49.3	49.3	50.5	69.4	67.0	64.3
ALBEF*+	48.2	48.4	49.7	75.3	72.7	69.5	59.2	57.3	56.0	70.6	68.3	65.5

our multi-granularity distillation strategy (denoted as "+") is combined with various methods. This phenomenon validates the effectiveness and generalization capability of our approach.

I. Exploration on Large-Scale Multi-modal Model

To explore the effectiveness of our proposed coordinate optimization strategy for large-scale multi-modal models, we apply our method to the fine-tuning process (marked with “+”). Table VII takes X-VLM as an example on the FLICKR30K dataset. Results demonstrate that our method not only improves the performance of single-modal retrieval in the fine-tuning stage of the pre-trained large-scale model, for both the weak and strong modalities but also maintains excellent performance on most metrics in cross-modal retrieval and mixed retrieval tasks.

J. Case Study

To analyze the retrieval visualizations, we randomly sample cross-modal and mixed retrieval cases from the FLICKR30K dataset to validate the effectiveness of our proposed method. The visualization examples are exhibited in Fig. 6 and Fig. 7, in which green ticks/boxes/values represent exactly aligned instances, red forks/boxes/values denote unaligned instances, the mixed retrieval instances are with ROUGE-L value (the larger the better). Considering the superior performance, we adopt our method here.

Fig. 6 shows the qualitative results of cross-modal retrieval using our method and VSRN. First, most of the retrieved cross-modal instances using our method are correct (shown as green ticks) on both the I2T and T2I retrieval. Some outputs are mismatched (shown as red forks), but reasonable, for example, (a) 5 contains similar semantic meanings to the image. On the other hand, our method outperforms the VSRN on both the I2T and T2I retrieval considering the same query. For example, (b) and (d) share the same image query, our method can retrieve the most aligned sentences, while VSRN fails, the reason that better structure-preserving can promote consistent representation learning in return.

Fig. 7 shows the qualitative results of mixed retrieval using our method compared with the SCAN. We find that our method can not only find accurate cross-modal instances but also find semantically similar intra-modal instances. For example, in the image query case (a), our method can retrieve similar images and exactly aligned cross-modal instances, while SCAN only retrieves images with lower similarities (i.e., lower ROUGE-L values), and several unaligned sentences. This phenomenon further validates the effectiveness of our proposed method in mixed retrieval. Moreover, we can intuitively find the differences between consistent representations learned by different cross-modal approaches in (d). Using the “harp player” case as an example, we can find the clustered instances in our method are all with “harp player” semantics, whereas the clustered instances in SCAN are with outliers.

V. CONCLUSION

In this paper, we find that the learned consistent representations from existing vision-language retrieval methods may affect single-modal retrieval performance. To explain this phenomenon, we identify the main cause as modal sufficiency, i.e., there exist weak and strong modalities, and hard cross-modal consistency may bring negative representation learning

to strong modality, leading to the destruction of instance structure. To address this problem, we develop a different way inspired by multi-task learning, which aims to learn two modal representations that can simultaneously ensure cross-modal consistency and single-modal structure. Extensive experiments on different datasets validate our method can achieve better single-modal retrieval accuracy whilst maintaining cross-modal retrieval capacity compared with the baselines. In future work, we aspire to theoretically elucidate the relationship between the degree of modality imbalance and retrieval performance. Additionally, addressing the challenges posed by modality imbalance in cross-modal retrieval tasks involving more than two modalities represents a significant area for further investigation.

REFERENCES

- [1] J. Lu, D. Batra, D. Parikh, and S. Lee, “Vilbert: Pre-training task-agnostic visiolinguistic representations for vision-and-language tasks,” in *NeurIPS*, 2019, pp. 13–23.
- [2] C. Huang, X. Luo, J. Zhang, Q. Liao, X. Wang, Z. L. Jiang, and S. Qi, “Explore instance similarity: An instance correlation based hashing method for multi-label cross-model retrieval,” *IPM*, vol. 57, no. 2, p. 102165, 2020.
- [3] J. Rao, L. Ding, S. Qi, M. Fang, Y. Liu, L. Shen, and D. Tao, “Dynamic contrastive distillation for image-text retrieval,” *TMM*, vol. 25, pp. 8383–8395, 2023.
- [4] F. Wan, X. Wu, Z. Guan, and Y. Yang, “Covlr: Coordinating cross-modal consistency and intra-modal relations for vision-language retrieval,” in *ICME*, 2024, pp. 1–6.
- [5] Y. Zhang, W. Zhou, M. Wang, Q. Tian, and H. Li, “Deep relation embedding for cross-modal retrieval,” *TIP*, vol. 30, pp. 617–627, 2021.
- [6] K. Lee, X. Chen, G. Hua, H. Hu, and X. He, “Stacked cross attention for image-text matching,” in *ECCV*, 2018, pp. 212–228.
- [7] P. Hu, Z. Huang, D. Peng, X. Wang, and X. Peng, “Cross-modal retrieval with partially mismatched pairs,” *TPAMI*, vol. 45, no. 8, pp. 9595–9610, 2023.
- [8] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, “VSE++: improving visual-semantic embeddings with hard negatives,” in *BMVC*, 2018, p. 12.
- [9] H. Diao, Y. Zhang, L. Ma, and H. Lu, “Similarity reasoning and filtration for image-text matching,” in *AAAI*, 2021, pp. 1218–1226.
- [10] Y. Peng, J. Qi, and Y. Yuan, “Modality-specific cross-modal similarity measurement with recurrent attention network,” *TIP*, vol. 27, no. 11, pp. 5585–5599, 2018.
- [11] H. Zhang, Z. Mao, K. Zhang, and Y. Zhang, “Show your faith: Cross-modal confidence-aware network for image-text matching,” in *AAAI*, 2022, pp. 3262–3270.
- [12] H. Diao, Y. Zhang, L. Ma, and H. Lu, “Similarity reasoning and filtration for image-text matching,” in *AAAI*, 2021, pp. 1218–1226.
- [13] G. Luo, Y. Zhou, X. Sun, Y. Wang, L. Cao, Y. Wu, F. Huang, and R. Ji, “Towards lightweight transformer

- via group-wise transformation for vision-and-language tasks,” *TIP*, vol. 31, pp. 3386–3398, 2022.
- [14] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, “BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation,” in *ICML*, 2022, pp. 12 888–12 900.
- [15] Z. Fu, Z. Mao, Y. Song, and Y. Zhang, “Learning semantic relationship among instances for image-text matching,” in *CVPR*, 2023, pp. 15 159–15 168.
- [16] W. Wang and Z. Zhou, “Co-training with insufficient views,” in *ACML*, 2013, pp. 467–482.
- [17] C. Liu, H. Ding, Y. Zhang, and X. Jiang, “Multi-modal mutual attention and iterative interaction for referring image segmentation,” *TIP*, vol. 32, pp. 3054–3065, 2023.
- [18] Y. Yang, H. Ye, D. Zhan, and Y. Jiang, “Auxiliary information regularized machine for multiple modality feature learning,” in *IJCAI*, 2015, pp. 1033–1039.
- [19] Y. Huang, J. Lin, C. Zhou, H. Yang, and L. Huang, “Modality competition: What makes joint training of multi-modal network fail in deep learning? (provably),” in *ICML*, vol. 162, 2022, pp. 9226–9259.
- [20] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” in *NAACL*, 2019, pp. 4171–4186.
- [21] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *ICCV*, 2021, pp. 9992–10 002.
- [22] L. Van der Maaten and G. Hinton, “Visualizing data using t-sne,” *JMLR*, vol. 9, no. 11, 2008.
- [23] Y. Yang, C. Zhang, Y. Xu, D. Yu, D. Zhan, and J. Yang, “Rethinking label-wise cross-modal retrieval from a semantic sharing perspective,” in *IJCAI*, 2021, pp. 3300–3306.
- [24] J. Yu, W. Zhang, Y. Lu, Z. Qin, Y. Hu, J. Tan, and Q. Wu, “Reasoning on the relation: Enhancing visual representation for visual question answering and cross-modal retrieval: Enhancing visual representation for visual question answering and cross-modal retrieval,” *TMM*, vol. 22, no. 12, pp. 3196–3209, 2020.
- [25] X. Xu, K. Lin, Y. Yang, A. Hanjalic, and H. T. Shen, “Joint feature synthesis and embedding: Adversarial cross-modal retrieval revisited,” *TPAMI*, vol. 44, no. 6, pp. 3030–3047, 2022.
- [26] X. Dong, L. Liu, L. Zhu, L. Nie, and H. Zhang, “Adversarial graph convolutional network for cross-modal retrieval,” *TCSV*, vol. 32, no. 3, pp. 1634–1645, 2022.
- [27] N. Messina, G. Amato, A. Esuli, F. Falchi, C. Gennaro, and S. Marchand-Maillet, “Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders,” *TOMCCAP*, vol. 17, no. 4, pp. 128:1–128:23, 2021.
- [28] 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 *ICML*, 2021, pp. 8748–8763.
- [29] P. Hu, H. Zhu, J. Lin, D. Peng, Y. Zhao, and X. Peng, “Unsupervised contrastive cross-modal hashing,” *TPAMI*, vol. 45, no. 3, pp. 3877–3889, 2023.
- [30] H. Tan and M. Bansal, “LXMERT: learning cross-modality encoder representations from transformers,” in *EMNLP*, 2019, pp. 5099–5110.
- [31] J. Li, R. R. Selvaraju, A. D. Gotmare, S. R. Joty, C. Xiong, and S. C. H. Hoi, “Align before fuse: Vision and language representation learning with momentum distillation,” in *NeurIPS*, 2021, pp. 9694–9705.
- [32] W. Wang, D. Tran, and M. Feiszli, “What makes training multi-modal classification networks hard?” in *CVPR*, 2020, pp. 12 695–12 705.
- [33] Z. Han, C. Zhang, H. Fu, and J. T. Zhou, “Trusted multi-view classification,” in *ICLR*, 2021.
- [34] X. Peng, Y. Wei, A. Deng, D. Wang, and D. Hu, “Balanced multimodal learning via on-the-fly gradient modulation,” in *CVPR*, 2022, pp. 8228–8237.
- [35] Y. Zeng, X. Zhang, and H. Li, “Multi-grained vision language pre-training: Aligning texts with visual concepts,” in *ICML*, 2022, pp. 25 994–26 009.
- [36] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *CVPR*, 2016, pp. 770–778.
- [37] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [38] 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,” in *ICLR*, 2021.
- [39] W. Park, D. Kim, Y. Lu, and M. Cho, “Relational knowledge distillation,” in *CVPR*, 2019, pp. 3967–3976.
- [40] S. Chen, G. Niu, C. Gong, J. Li, J. Yang, and M. Sugiyama, “Large-margin contrastive learning with distance polarization regularizer,” in *ICML*, 2021, pp. 1673–1683.
- [41] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H. Shum, “DINO: DETR with improved denoising anchor boxes for end-to-end object detection,” in *ICLR*, 2023.
- [42] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *JMLR*, vol. 21, pp. 140:1–140:67, 2020.
- [43] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *ECCV*, 2014, pp. 740–755.
- [44] A. Karpathy and F. Li, “Deep visual-semantic alignments for generating image descriptions,” in *CVPR*, 2015, pp. 3128–3137.
- [45] M. J. Huiskes and M. S. Lew, “The MIR flickr retrieval evaluation,” in *ACMMM*, 2008, pp. 39–43.
- [46] D. Gurari, Y. Zhao, M. Zhang, and N. Bhattacharya, “Captioning images taken by people who are blind,” in *ECCV*, 2020, pp. 417–434.
- [47] H. Chen, G. Ding, X. Liu, Z. Lin, J. Liu, and J. Han,

- “IMRAM: iterative matching with recurrent attention memory for cross-modal image-text retrieval,” in *CVPR*, 2020, pp. 12 652–12 660.
- [48] C. Liu, Z. Mao, T. Zhang, H. Xie, B. Wang, and Y. Zhang, “Graph structured network for image-text matching,” in *CVPR*, 2020, pp. 10 918–10 927.
- [49] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, “Visual semantic reasoning for image-text matching,” in *ICCV*, 2019, pp. 4654–4662.
- [50] K. Zhang, Z. Mao, Q. Wang, and Y. Zhang, “Negative-aware attention framework for image-text matching,” in *CVPR*, 2022, pp. 15 640–15 649.
- [51] S. Goel, H. Bansal, S. Bhatia, R. A. Rossi, V. Vinay, and A. Grover, “Cyclip: Cyclic contrastive language-image pretraining,” in *NeurIPS*, 2022, pp. 6704–6719.
- [52] C. Du, T. Li, Y. Liu, Z. Wen, T. Hua, Y. Wang, and H. Zhao, “Improving multi-modal learning with uni-modal teachers,” *CoRR*, vol. abs/2106.11059, 2023.
- [53] J. Tang, Z. Li, M. Wang, and R. Zhao, “Neighborhood discriminant hashing for large-scale image retrieval,” *TIP*, vol. 24, no. 9, pp. 2827–2840, 2015.
- [54] F. Wang, J. Pan, S. Xu, and J. Tang, “Learning discriminative cross-modality features for rgb-d saliency detection,” *TIP*, vol. 31, pp. 1285–1297, 2022.
- [55] Y. Gao, J. Tang, R. Hong, S. Yan, Q. Dai, N. Zhang, and T. Chua, “Camera constraint-free view-based 3-d object retrieval,” *TIP*, vol. 21, no. 4, pp. 2269–2281, 2012.
- [56] F. Carrara, A. Esuli, T. Fagni, F. Falchi, and A. M. Fernández, “Picture it in your mind: generating high-level visual representations from textual descriptions,” *Information Retrieval Journal*, vol. 21, no. 2-3, pp. 208–229, 2018.
- [57] L. Huang, W. Wang, J. Chen, and X. Wei, “Attention on attention for image captioning,” in *ICCV*, 2019, pp. 4633–4642.
- [58] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *ICLR*, 2019.