GMA: Green Multi-Modal Alignment for Image-Text Retrieval

Tsung-Shan Yang Yun-Cheng Wang Chengwei Wei Suya You C.-C. Jay Kuo
University of Southern California, USA
{tsungsha, yunchenw, chengwei, suya, jckuo}@usc.edu

***Abstract*—Image-text retrieval is a fundamental task in image understanding. This task aims to retrieve the most relevant information from another modality based on the given image or text. Recent approaches focus on training large neural networks to bridge the gap between visual and textual domains. However, these models are computationally expensive and not explainable regarding how the data from different modalities are aligned. End-to-end optimized models, such as large neural networks, can only output the final results, making it difficult for humans to understand the reasoning behind the model's predictions. Hence, we propose a green learning solution, Green Multi-Modal Alignment (GMA), for computational efficiency and mathematical transparency. We reduce trainable parameters to 3% compared to fine-tuning the whole image and text encoders. Experimental results show that our model can outperform the SOTA retrieval models in text-to-image and image-to-text retrieval on the Flick30k and MS-COCO datasets. Besides, our alignment process can incorporate visual and text encoder models trained separately and generalize well to unseen image-text pairs.**

***Index Terms*—Image-text retrieval, Multimodal Alignment, Green Learning, Image Understanding.**

I. Introduction

Image-text retrieval is a fundamental step of image understanding in computer vision. Real-world information can be represented as an array of pixels or textual descriptions. Texts and images are the interfaces to access stored knowledge. The image-text retrieval task aims to find the most relevant counterparts of the image-text pairs given either an image or a textual description. Fig. 1 shows an example of an image and its paired textual descriptions.

1. [x] A damaged building has a an excavator in front of it.
2. [x] A bulldozer works to demolish a decrepit building; in the background, another brick building waits for its demise, its face covered with a grid of blackened window-holes.
3. [x] Construction equipment at work.
4. [x] Heavy machinery in a construction zone.
5. [x] A crane operates amidst piles of rubble.
6. [x] A yellow construction vehicle is posed near two buildings, its arm engaged with a pile of rubble.
7. [HIT] A man in a yellow coat is on a blue industrial crane working on the side of a tall building.
8. [HIT] A worker in a yellow jacket is hoisted up high to work on a building.
9. [x] A man using a bulldozer on a construction site.
10. [HIT] A Heavy machine lifting up a worker.

Fig. 1. The example of image-to-text retrieval. By giving an image, we need to retrieve the paired captions from the candidate set.

One of the challenges in image-text retrieval is explainability. Humans will expect a complete reasoning procedure instead of a magic answer from the model. However, the complicated models hide the reasoning procedures in the numerical latent spaces, and the nonlinearities in the model make the whole inference process a black box. To this end, we propose a stagewise methodology, dividing the retrieval process into three stages: 1) Global Alignment, 2) Image Cluster Alignment, and 3) Text Cluster Alignment. Each alignment stage consists of three modules: a) alignment, b) subdomain clustering, and c) subdomain feature selection. More fine-grained information can be revealed in the module's feature selection process.

The second issue is the availability of paired image and text data. It is critical to have high-quality pairs in both domains when we train the dual encoders jointly. However, most datasets only contain high-quality data in a single modality. For example, ImageNet [1] and MS-COCO [2] contain diverse images but lack sentence-level textual descriptions associated with the images. On the contrary, in the textual datasets, the BooksCorpus (800M words) [3] and English Wikipedia (2,500M words) contain high-quality paragraphs yet without corresponding images. Instead of jointly training text and image encoders from scratch, we adopt the pre-trained encoders in the image and text domains. Then, we proposed a green-learning alignment process to deal with the lack of paired information.

We propose a new scheme called Green Multi-Modal Alignment (GMA). The method utilizes the frozen image and text encoder models and aligns the representations using the proposed alignment process. Our contributions are summarized as follows:

- Instead of fine-tuning the pre-trained encoders, we design the stagewise alignment procedure. The number of trainable parameters is around 3% compared to fine-tuning the whole encoders, making our model computationally efficient.
- The modularized design provides explainability in the retrieval process. The task can be divided into subproblems. Since the token importance can defined in the clustering and feature selection modules, we can understand the alignment statistically within the stages.
- The stagewise alignments are linear projections without any nonlinearity. Thus, the alignment process can be easily reversed from one to another.
- We conduct extensive experiments on two public multi-

modal datasets. The results demonstrate that our method can significantly improve the performance in text-to-image retrieval.

II. Related Work

The existing methods can be divided into two categories: 1) Cross-Modal retrieval and 2) Visual-Language models (VLMs). Cross-modal models consist of Convolutional Neural Networks (CNNs) for extracting features from images and Recurrent Neural Networks (RNNs) for processing text data. On the other hand, VLMs employ Large Language Models (LLMs) that work in tandem with the visual foundational models for optimal performance.

A. Cross-Modal Retrieval

The cross-modal retrieval algorithms consist of feature extractors and representation matching. Zheng et al. [4] adopt deep CNN as the backbone to extract the image features and deep RNN as the backbone to obtain text features. The instance loss optimizes the two feature extractors, which can project the representations from different modalities onto the joint latent space. Lee et al. [5] utilize bottom-up attention object detector [6] to obtain semantic representations of images and conduct the word-level matching in the captions.

Liu et al. [7] formulate the information as a graph and adopt the structural matching to retrieve the closest sub-graph. The finer image features can be obtained from the region of interest, and the finer word representations can be formulated from the part-of-speech(pos) tagging [8].

As the metric learning, Hadsell et al. [9] propose the idea of contrastive learning. The objective function aims to increase the distance between unpaired image and text representations while reducing the distance between paired representations. However, pairwise optimization relies heavily on the quality of paired data. We align two pre-trained encoders trained separately in the text and image domains to overcome the need for paired data.

B. Visual-Language Model

Transformers [10] have significantly succeeded in natural language processing and computer vision tasks. The image-text encoders can share similar architectures. CLIP [11] demonstrates the impressive visual representations jointly trained with the paired text descriptions. The model uses a contrastive learning scheme to project image and text representations onto a shared latent space. This shared space allows for a better understanding of the relationship between the two modalities. The dual-encoder(image-text) architecture is prevalent in multimodal applications.

Despite achieving state-of-the-art performance, large visual-language pre-trained models still have shortcomings in inference. The matching process is not transparent, and humans can't understand the decision-making within the fully connected layers as they lack semantic meanings. Apart from explainability, the jointly fine-tuning process is computationally expensive. To handle the transparency and efficiency, we introduce the Green Learning Alignment algorithm, which uses separately pre-trained image-text encoders. The idea of Green Learning was proposed by Kuo et al. [12] and aims to reduce the computational cost of backpropagation while providing a theoretically explainable learning process for various applications.

III. Method

The algorithm can be divided into three stages: 1) Global Alignment, 2) Image Cluster Alignment, and 3) Text Cluster Alignment. We adopt the stage-wise approach to approximate the complicated decision-making process rather than building the visual-language fundamental models from scratch. Starting from the pre-trained image and text feature extractors, we keep the pre-trained model frozen to maintain its ability to generalize with unpaired data in the matching process. We align the representations by training additional one-layered adaptor matrices to project the representations onto the joint latent space. Precisely, the alignment process consists of three modules: a) alignment, b) sub-domain clustering, and c) sub-domain feature selection, where sub-domain clustering and feature selection are conducted in both image and text domains, as shown in Fig.2.

A. Alignment

In the alignment process, we do not fine-tune the pre-trained encoders. We train a lightweight linear transformation in the visual and textual domains to align the two representation spaces. The visual and text embeddings can be formulated as:

$$\begin{aligned} e_{vis} &= \mathcal{F}(\text{Image}) \in \mathbf{R}^{d_{vis}} \\ e_{txt} &= \mathcal{G}(\text{Caption}) \in \mathbf{R}^{d_{txt}}, \end{aligned} \tag{1}$$

where e_{vis}, e_{txt} are the image and text embeddings, $\mathcal{F}, \mathcal{G}$ are the frozen image and text encoder models, and d_{vis}, d_{txt} are the dimensions of the image and text representations. With the deterministic representations, the matching process can be denoted as:

$$sim(Ae_{vis}, Be_{txt}) = sim(z_{vis}, z_{txt}), \tag{2}$$

where $A \in \mathbf{R}^{d_{joint} \times d_{vis}}$ and $B \in \mathbf{R}^{d_{joint} \times d_{txt}}$ represent the trainable image-text alignment matrices, $z \in \mathbf{R}^{d_{joint}}$ represents the vector in the joint space, and $sim(.,.)$ represents the similarity metric. We adopt the cosine similarity as the similarity metric, namely $sim(u, v) = \frac{u \cdot v}{\|u\|\|v\|}$. We can further optimize the trainable parameters with the contrastive learning loss function [13].

$$\mathcal{L}_{con} = -\log \frac{exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{N} \mathbb{1}_{[k \neq i]} exp(sim(z_i, z_k)/\tau)}. \tag{3}$$

Here, (i, j) denotes the paired image and sentence in the sampled batch, N denotes the batch size, and $\tau \in \mathbf{R}$ denotes the temperature hyperparameter. $\mathbb{1} \in \{0, 1\}$ is an indicator function, and the value is one while $[k \neq i]$. The objective function collects the representations of paired samples and pushes apart the distances of unpaired samples in the latent space.

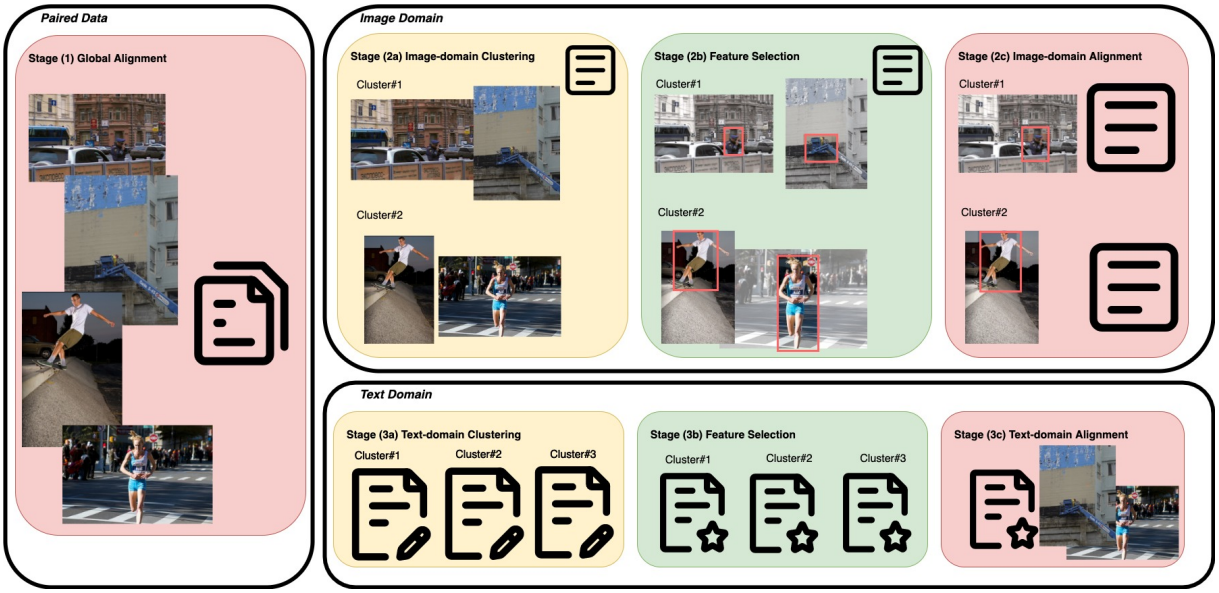

Fig. 2. The overall algorithm design of Alignment. The first stage is the global alignment. The second and third stages include fine-grained clustering and image- and text-domain feature selections.

Hence, the problem can be formulated as an optimization problem, and all the transformations are linear. We can define the inverse projection in the joint latent space without nonlinearities.

$$\begin{aligned} e_{vis} &= A^{-1} z_{vis} \\ e_{txt} &= B^{-1} z_{txt}, \end{aligned} \tag{4}$$

where $A^{-1} \in \mathbf{R}^{d_{vis} * d_{joint}}$ and $B^{-1} \in \mathbf{R}^{d_{txt} * d_{joint}}$ represent the inverse transformation from the joint space to the original image-text representations. We define the reconstruction loss for both the image and text modality.

$$\begin{aligned} \mathcal{L}_{recon} = &\left\| A^{-1} z_{vis} - e_{vis} \right\|_2 \\ &+ \left\| B^{-1} z_{txt} - e_{txt} \right\|_2, \end{aligned} \tag{5}$$

The alignment process is a linear transformation conducted by the A and B matrices. The objective function can be denoted as:

$$\mathcal{L} = \alpha \mathcal{L}_{con} + \beta \mathcal{L}_{recon}, \tag{6}$$

where α, β, and $\gamma \in \mathbf{R}$ represent hyperparameters in training. The linear alignment provides the invertible transformation from the image-text modality to the joint latent space and vice versa. However, the single-layered alignment is too simple to match all the samples. Thus, we cluster the data to form subdatasets and utilize the stagewise alignments for the detailed decision.

B. Sub-domain Clustering

With the alignment process, we can find similar representations by linear transformations. However, the transformation can only take the global representations, which means that the images/captions are represented as d_{vis}-/d_{txt}-dimensional vectors. In previous research, fine-grained information was also crucial in information-matching tasks. The image/sentence representations are the pooled output from the tokens in the prevailing transformer models.

Due to the complexity of the fine-grained token representations, it is challenging to train the tokenwise alignment in a brutal force manner. Thus, we adopt the clustering algorithms and use the clustering results to obtain the crucial tokens. The crucial token selection will be introduced in section III-C. We can reduce the feature dimension from the number of tokens and perform a second-stage alignment.

The clustering is based on the KMeans algorithm. To ensure the consistency of alignment and clustering, we use the l_2-norm of normalized representations as the distance metric.

$$\|\tilde{u} - \tilde{v}\|_2^2 = \|\tilde{u}\|_2^2 + \|\tilde{v}\|_2^2 - 2\tilde{u}\tilde{v} = 2 - 2sim(u, v), \tag{7}$$

where $\tilde{u}$ and $\tilde{v}$ are the normalized representations, namely $\tilde{u} = \frac{u}{\|u\|}$. The clustering probability can be denoted as

$$\begin{aligned} prob(u \in clus_i) &= \frac{e^{\epsilon'(2-2sim(u,cen_j))}}{\sum_{j=1}^{K} e^{\epsilon'(2-2sim(u,cen_i))}} \\ &= \frac{e^{\epsilon \cdot sim(u,cen_i)}}{\sum_{j=1}^{K} e^{\epsilon \cdot sim(u,cen_j)}}, \end{aligned} \tag{8}$$

where $clus_i$ and cen_i represent the i-th cluster and i-th centroid vector, respectively. K represents the number of clusters and ϵ is a hyperparameter. If the ϵ increases, the probability distribution will concentrate on a certain class. If the ϵ decreases, the probability distribution will become uniform.

C. Feature Selection

Clustering results provide pseudo-labels for further feature selections. We adopt Discriminant Feature Selection [14] (DFT) to select informative and reduce feature dimensions.

DFT is a supervised feature selection process that measures the dimension-wise importance. For a given 1D input feature, we can order the samples by the feature values and bind the feature dimension with the sample maximum and sample minimum. Then, we can partition the samples along the given dimension and calculate the partition purity by weighted cross-entropies with the pseudo-labels obtained from section III-B. A feature is more discriminant if it has a lower loss value. Then, we can plot the loss value curve from the lowest to the highest and use the elbow point to select discriminant features from the whole feature set.

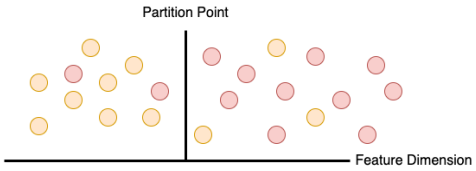

Fig. 3. Visualization of DFT. Red and orange dots represent the binary labels. The partition metric is the weighted sum of the left and right binary cross-entropy.

Separating the whole dataset into subsets allows us to conduct the discriminant feature test among the tokens with the pseudo-labels from the clustering results. Thus, the token-level alignments can be performed using the same procedure as the global-level alignment.

D. Mathematical Expression

The overall alignment process can be divided into three modules: 1) global matching, 2) sub-domain clustering, and 3) sub-domain matching. The sub-domain clustering and alignment will be conducted within the image and the text domain. We can aggregate the alignments in the sub-problem to approximate the overall alignment.

$$\begin{aligned} &sim(image, text) \\ &= sim(W^{vis}\mathcal{F}(\text{image}), W^{txt}\mathcal{G}(\text{text})) \\ &= sim(W^{vis}[G_{vis}; T_{vis}], W^{txt}[G_{txt}; T_{txt}]) \\ &\sim sim(W^{vis}_{global}G_{vis}, W^{txt}_{global}G_{txt}) \\ &+ sim(W^{vis}_{tokens}[G_{vis}; T_{vis}], W^{txt}_{g}G_{txt}) \\ &+ sim(W^{vis}_{global}G_{vis}, W^{txt}_{tokens}[G_{txt}; T_{txt}]), \end{aligned} \tag{9}$$

where G_{vis} and G_{txt} denote the pooled outputs from the feature extractors (global features), T_{vis} and T_{txt} denote the token, i.e. fine-grained, features, $W^{\cdot}_{\cdot}$ denote the alignment matrices corresponding to different subsets from the clustering results. Due to the computational cost, we cannot directly collect all token features. Therefore, we conduct the feature selection process based on the clustering results.

The feature selection process is an approximation based on the clustering results. The process is denoted as the combination of the conditional probabilities. For simplicity, we ignore the alignment matrix in the following representations.

$$\begin{aligned} &E[sim(\mathcal{F}(\text{image}), \mathcal{G}(\text{text}))] \\ &= E[E[sim(\mathcal{F}(\text{image}), \mathcal{G}(\text{text}))]|\text{image} \in C_1; \text{text} \in C_2] \\ &\sim E[sim(G_{vis}, G_{txt})] \\ &+ E[E[sim(DFT([G_{vis}; T_{vis}]), G_{txt})|\text{image} \in C_1]] \\ &+ E[E[sim(G_{vis}, DFT([G_{txt}; T_{txt}]))|\text{text} \in C_2]], \end{aligned} \tag{10}$$

where $DFT(.)$ represents the feature selection and dimension reduction process in section III-C, and C_1 and C_2 represent the KMeans cluster sets. Instead of training a complicated alignment process from the token level output of the feature extractor, we propose the stagewise decomposition on the dataset and train simpler structures for the subsets. Meanwhile, the alignments in the stages are linear, which provides the inversion operation and preserves the dual accessibility in both image and text domains.

IV. Experiments

A. Dataset

We conduct the image-to-text and text-to-image retrieval on the image-text benchmark: Flickr30k and MS-COCO. The Flickr30k dataset [15] contains 31,000 images, and every image has five paired captions. The training set contains 29,000 images; the validation and testing sets contain 1,000. The MS-COCO [2] is a larger-scale dataset with 123,287 images, each containing at least five captions. We follow the 'Karpathy' splitting for the experiments [16]: 113,287 images for training, 5,000 for validation, and 5,000 for testing. We use the two benchmarks with different sizes to demonstrate the scalability and generalizability of our approaches. The performance is evaluated using the Recall@K metric where $K \in \{1, 5, 10\}$. The notation K denotes the top-K matches from the candidate set. The retrieval will be considered a true positive once the predicted matches include the paired ground truth.

B. Retrieval

We conducted the experiments and compared our alignment approach to the SOTA retrieval models. The results are shown in Table. I. In the experiments, we extract information from the frozen CLIP image and text encoder. The CLIP encoder contains more than 428M parameters. However, we do not fine-tune the overall encoder in our alignment process; instead, we train additional alignment matrices. The trainable parameters can be reduced from 428M to 9.43M (2.2%).

In Flickr30k (1k testing set), our approach outperforms other image-to-text and text-to-image retrieval methods. The alignment can improve the recall@1 by 0.6% in the image-to-text retrieval. Meanwhile, our approach provides a 6% boost in text-to-image retrieval. RCAR [17] needs dual-way optimized models, namely image-to-text and text-to-image. Our method is optimized in a feed-forward manner and ensembles the substructures directly.

In MS-COCO (5k testing set), our method provides competitive performance in image-to-text retrieval and outperforms

TABLE I
THE FLICKR30K(1K TESTING SET) AND MSCOCO(5K TESTING SET) DATASET RETRIEVAL PERFORMANCE. WE COMPARE THE SINGLE-MODEL PERFORMANCE AMONG ALL MULTI-MODAL RETRIEVAL MODELS. THE NUMBERS ARE TAKEN FROM DIAO ET AL. [17] R@1 REPRESENTS RECALL@1 FOR SIMPLICITY.

	Flickr30k (1k testing set)						MS-COCO (5k testing set)					
	image-to-text			text-to-image			image-to-text			text-to-image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
SCAN [5]	67.4	90.3	95.8	48.6	77.7	85.2	50.4	82.2	90.0	38.6	69.3	80.4
VSRN [18]	71.3	90.6	96.0	54.7	81.8	88.2	53.0	81.1	89.4	40.5	70.6	81.1
CAAN [19]	70.1	91.6	97.2	52.8	79.0	87.9	52.5	83.3	90.9	41.2	70.3	82.9
IMRAM [20]	74.1	93.0	96.6	53.9	79.4	87.2	53.7	83.2	91.0	39.7	69.1	79.8
MMCA [21]	74.2	92.8	96.4	54.8	81.4	87.8	54.0	82.5	90.7	38.7	69.7	80.8
GSMN [7]	76.4	94.3	97.3	57.4	82.3	89.0	–	–	–	–	–	–
SGRAF [22]	77.8	94.1	97.4	58.5	83.0	88.8	57.8	84.9	91.6	41.9	70.7	81.3
SHAN [23]	74.6	93.5	96.9	55.3	81.3	88.4	–	–	–	–	–	–
WCGL [24]	74.8	93.3	96.8	54.8	80.6	87.5	–	–	–	–	–	–
RCAR [17]	78.7	94.6	97.6	59.5	84.0	89.5	**59.6**	**85.8**	**92.4**	42.5	71.7	81.8
SGRAFS [8]	79.2	95.3	97.7	58.3	83.1	89.2	58.0	85.1	91.6	41.7	71.2	81.5
CLIP [11]	88.0	98.7	99.4	68.7	90.6	95.2	58.4	81.5	88.1	37.8	62.4	72.2
GMA(Ours)	**88.6**	**98.9**	**99.6**	**74.8**	**93.5**	**96.7**	58.6	83.2	90.0	**45.3**	**72.6**	**82.8**

the others in text-to-image retrieval by a boost of 2.1% in Recall@1. We achieve the best text-to-image retrieval performance among the two datasets, showcasing our approach's scalability.

C. Visual Feature Comparison

This section demonstrates the alignment between the visual/text encoders, which are trained separately. Starting from the jointly trained CLIP structure, we change the text encoders into the RoBERTa [25] and the visual encoder into a CNN-based object detector [6].

The results are shown in Table II. The best performance comes from the jointly train models, whose representations are preliminarily aligned in the pre-trained process. Compared to the CLIP visual encoder, the features from the object detector are weaker in the alignment process. On the other hand, the separately trained text encoder, RoBERTa [25], does not take harm from the unpaired training dataset. The representations from the CLIP visual encoder and the RoBERTa text encoder can provide a competitive performance in image-to-text retrieval and a better performance in text-to-image retrieval than the original CLIP. The encoder can be adapted to the retrieval application without fine-tuning with the paired text/text data.

V. CONCLUSION AND FUTURE WORK

Our approach can achieve outstanding performance in both image-to-text and text-to-image retrieval tasks. Furthermore, our method involves a step-by-step alignment process that maintains compatibility in the decision-making procedure. We divide the alignment into global and sub-domain matching and apply a feature selection method to decrease the input feature dimensions. All the sub-processes can be expressed statistically rather than the black-box outputs. To ensure computational efficiency, we have frozen the visual and text encoders and only trained the alignment matrices, accounting for only about 3% of the parameters compared to the original model.

We have tested our approach of aligning visual and text encoders trained separately. In the testing dataset, we found that the pre-trained text encoder can improve the performance of text-to-image retrieval. Replacing the text encoder can also lead to similar performance in image-to-text retrieval.

We are working on developing a purely green learning solution to image understanding in the visible future. By aiming not only for transparency but also computational efficiency, we can have a better understanding of the multimodal information representation.

REFERENCES

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[2] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, Springer, 2014, pp. 740–755.

[3] Y. Zhu, R. Kiros, R. Zemel, *et al.*, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 19–27.

[4] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, and Y.-D. Shen, "Dual-path convolutional image-text embeddings with instance loss," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 2, pp. 1–23, 2020.

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

TABLE II
THE EXPERIMENT RESULTS WITH DIFFERENT VISUAL AND TEXT FEATURES FOR THE ALIGNMENT PROCESS. ALL THE EXPERIMENTS ARE CONDUCTED IN THE FLICKR30K DATASET.

Visual *enc.*	Text *enc.*	Alignment (GMA)	Flickr30k (1k testing set)					
			image-to-text			text-to-image		
			Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10
CLIP vis [11]	CLIP text [11]	x	<u>88.0</u>	<u>98.7</u>	<u>99.4</u>	68.7	90.6	95.2
Object detector [26]	RoBERTa [25]	v	36.0	65.3	77.6	53.2	69.1	70.2
Object detector [26]	CLIP text [11]	v	33.2	62.7	75.4	49.8	65.1	68.2
CLIP vis [11]	RoBERTa [25]	v	86.3	98.2	<u>99.4</u>	<u>73.2</u>	**93.5**	**96.9**
CLIP vis [11]	CLIP text [11]	v	**88.6**	**98.9**	**99.6**	**74.8**	**93.5**	<u>96.7</u>

[6] P. Anderson, X. He, C. Buehler, *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.

[7] C. Liu, Z. Mao, T. Zhang, H. Xie, B. Wang, and Y. Zhang, "Graph structured network for image-text matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 921–10 930.

[8] B. Jawade, D. D. Mohan, N. M. Ali, S. Setlur, and V. Govindaraju, "Napreg: Nouns as proxies regularization for semantically aware cross-modal embeddings," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 1135–1144.

[9] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, IEEE, vol. 2, 2006, pp. 1735–1742.

[10] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[11] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, PMLR, 2021, pp. 8748–8763.

[12] C.-C. J. Kuo and A. M. Madni, "Green learning: Introduction, examples and outlook," *Journal of Visual Communication and Image Representation*, vol. 90, p. 103 685, 2023.

[13] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, PMLR, 2020, pp. 1597–1607.

[14] Y. Yang, W. Wang, H. Fu, C.-C. J. Kuo, *et al.*, "On supervised feature selection from high dimensional feature spaces," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

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

[16] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

[17] H. Diao, Y. Zhang, W. Liu, X. Ruan, and H. Lu, "Plug-and-play regulators for image-text matching," *IEEE Transactions on Image Processing*, 2023.

[18] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Visual semantic reasoning for image-text matching," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4654–4662.

[19] Q. Zhang, Z. Lei, Z. Zhang, and S. Z. Li, "Context-aware attention network for image-text retrieval," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3536–3545.

[20] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 655–12 663.

[21] X. Wei, T. Zhang, Y. Li, Y. Zhang, and F. Wu, "Multi-modality cross attention network for image and sentence matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 941–10 950.

[22] H. Diao, Y. Zhang, L. Ma, and H. Lu, "Similarity reasoning and filtration for image-text matching," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, 2021, pp. 1218–1226.

[23] Z. Ji, K. Chen, and H. Wang, "Step-wise hierarchical alignment network for image-text matching," *arXiv preprint arXiv:2106.06509*, 2021.

[24] Y. Wang, T. Zhang, X. Zhang, *et al.*, "Wasserstein coupled graph learning for cross-modal retrieval," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, 2021, pp. 1793–1802.

[25] Y. Liu, M. Ott, N. Goyal, *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[26] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.