APSIPA Transactions on Signal and Information Processing, XXXX, XX: 1–30

Original Paper

Image-Text Retrieval via Green Explainable Multi-modal Alignment (GEMMA)

Tsung-Shan Yang[1], Yun-Cheng Wang[1], Chengwei Wei[1], Suya You[2] and C.-C. Jay Kuo[1]

[1] *University of Southern California, Los Angeles, California, USA*
[2] *DEVCOM Army Research Laboratory, Adelphi, Maryland, USA*

ABSTRACT

Image-text retrieval is a fundamental task in image understanding. The algorithm fetches the most relevant counterpart in the other modality by giving the image or text. Large visual-language models are trained by paired image and text data to extract the joint representations. However, they are computationally expensive and not explainable regarding how the data from different modalities are aligned. To this end, we propose an efficient and stage-wise alignment for image and text representations, called the Green Explainable Multi-Modal Alignment (GEMMA). GEMMA is computationally efficient by reducing trainable parameters to 3% compared to fine-tuning all image and text encoders. The intermediate clustering results demonstrate the explainability of the alignment mechanism in our model. Experiments show that GEMMA outperforms state-of-the-art retrieval models in text-to-image and image-to-text retrieval tasks on the Flick30k and MS-COCO datasets. GEMMA

*Corresponding author: tsungsha@usc.edu.

ISSN 2048-7703; DOI 10.1561/116.xxxxxxxx

can also be generalized to unseen image-text pairs from pre-trained visual and text encoders separately.

Keywords: Image-text retrieval, Multi-modal Alignment, Green Learning, Image Understanding

1 Introduction

Image-text retrieval links textual and visual information and is a foundational image understanding application in computer vision. The goal of the task is to link textual descriptions and pixels in image arrays that represent similar concepts or semantics. The image-text retrieval task aims to find the most relevant information from the candidate sets in the counterpart modality. That is, when an image is given, the model needs to extract related captions by ranking them with higher scores and vice versa. Fig. 1 shows an example of an image and its paired textual descriptions.

Image-text retrieval can provide the information for visual-textual applications, including visual question answering [27], image captioning [1], visual grounding [38], and visual common sense reasoning [48]. With the thriving development of deep learning and computational resources, neural networks dominate the current research trend. Jointly trained neural network-based image and text encoders transform the input text and image into vectors in a common latent space. The two encoders are trained under metric learning schemes, which compare the cosine similarity between the paired and unpaired image and text samples. For example, an intuitive solution to representing the image and text in a joint latent space is optimizing two encoder models by minimizing contrastive loss [4]. The loss function can gather the paired information but repel the unpaired data in the latent space.

Although end-to-end solutions perform astonishingly, explainability is crucial for image-understanding applications. In the multi-modal application scenario, humans expect a complete reasoning procedure instead of a magic answer from the model. However, neural networks obscure the reasoning process within the joint latent space through complex floating-point operations, e.g., calculating cosine similarities

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.

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

between vectors. The nonlinearities in the model make the whole inference process a black box. To this end, we propose a multi-stage 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 is revealed in the module's feature selection process.

The availability of paired image and text data is another challenge when training multi-modal models. Most datasets contain only high-quality data in a single modality. For example, ImageNet [7] and MS-COCO [23] contain diverse images but lack sentence-level textual descriptions associated with the images. In contrast, in textual datasets, the BooksCorpus (800M words) [52] and English Wikipedia (2,500M words) contain well-structured paragraphs, yet without corresponding images. Collecting paired images and captions is expensive and labor-intensive. Due to the subjectiveness of caption labels, it is impractical to assume consistent captions for one image. However, the quality of

collected pairs in both domains significantly impacts the performance of the jointly trained multi-modal encoders. Aiming to relieve the data scarcity, we adopt the pre-trained encoders in the image and text domains instead of jointly training text and image encoders from scratch. Then, we proposed a green learning alignment process to deal with the lack of paired information.

We propose a new Green Explainable Multi-Modal Alignment (GEMMA) scheme to deal with paired data scarcity and explainability. 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:

- We reduce the number of parameters to around 3% compared to fine-tuning the whole encoders. Instead of fine-tuning the pre-trained encoders, we propose an alignment scheme from two pre-trained encoders, making the pipeline computationally efficient.

- In order to achieve pipeline transparency, we narrow the set of candidates in a stage-wise manner. The modular design divides the entire dataset into subsets. We can statistically understand the retrieval process and the crucial tokens by the feature selection modules in the sub-domain clustering.

- We provide bidirectional retrieval in the proposed pipeline. The alignment modules consist of linear projections without incorporating 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.

2 Related Work

The existing methods can be classified into 1) cross-modal retrieval and 2) visual-language models (VLMs). Cross-modal models consist of a convolutional neural network (CNN) to extract features from images and a recurrent neural network (RNN) to process text data. The joint

representations of the convolution and recurrent backbones are optimized by metric learning. On the other hand, VLMs employ Large Language Models (LLMs) that work in tandem with the Visual Transformer models (ViTs) for optimal performance. The VLM optimization can be performed by contrastive learning, masking filling, and generative matching.

2.1 Cross-Modal Retrieval

The cross-modal retrieval algorithms consist of representation matching and feature extraction. Metric learning schemes measure the similarity between the samples and predict the matching scores. Hadsell *et al.* [13] propose the idea of contrastive learning. The loss formulation aims to reduce the distance in the latent space for similar samples and to increase the distance for different samples. Triplet loss [32], lifted structure loss [28], and N-Pair loss [35] construct the joint latent space by sampling training data. The losses gather the positive and repel the negative sampling schemes from positive and negative pairs, forming the positive and negative pairs with the sampling schemes. Thus, optimization can be improved by the hard sampling process [31, 43]. With the thriving development of self-supervised applications, SimSCE [12] and SimCLR [4] provide metrics to reinforce the representations. The losses map the origin and representations from the augmented images (crop, rotate, color distort, etc.) onto the same latent space.

Frome *et al.* [11] first proposed the concept of joint image and text embedding in the ImageNet [7] classification. The pipeline utilizes the textual information from the label to construct a lookup table from the nearby concepts as the target embedding, leading to a hierarchical classification. Zheng *et al.* [51] adopts deep CNN as the basis for extracting the image and 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.* [21] utilize bottom-up attention object detector [1] to obtain semantic representations of images and to perform word-level matching in the captions. The bottom-up detector can provide the modifiers with the noun, matching the corresponding sentence with the details.

C. Liu *et al.* [24] formulate the information as a graph and adopt the structural matching to retrieve the closest subgraph. The object

detector obtains the visual graph. The node features are the region of interest (ROI) feature of the model, and the vertices are constructed by the Multi-Layer Perceptron (MLP). The textual graph is the Part-Of-Speech (POS) prediction from the Gated Recurrent Unit (GRU) Networks. L. Wang *et al.* [37] adopts the instance-wise matching for the subgraphs. The overall matching score aggregates the partial graph similarities in a bottom-up manner.

To further exploit the information in the query image, Cheng *et al.* [6] adopts the optical character recognition (OCR) module to extract semantic information such as text embeddings of the scene. The model fuses the image token and the scene text for the joint representation. Diao *et al.* [9] build the image tokens from ROI by the object detector and bidirectional GRU textual tokens. The cross-modal attention module is used for the token-wise matching process. Jawade *et al.* [15] constructs the visual and textual tokens from the pre-trained model. However, the research merges the cross-modal information by cross-attention [36] modules and manages the retrieval task with the transformer structures.

2.2 Visual-Language Model

Transformers [36] have achieved significant results in natural language processing and computer vision tasks. The image-text encoders can share similar architectures. W. Wang *et al.* [39] crop the input images into patches and use the patches as visual tokens to formulate the images as a novel language. The jointly trained visual and text encoders [5] [49] are optimized end-to-end. Visual language models (VLMs) can be categorized into three families [29] by the optimization process: (1) contrast-based VLMs, (2) VLMs with masking objects, and (3) generative-based VLMs. Constructive VLMs [30] are trained by the paired multi-modal data, and the objective loss is the contrastive loss. The self-supervised learning scheme obtains VLMs with masking objects [18, 20, 33]; the model needs to predict the masked visual and textual tokens. Generative-based VLMs [46, 47, 25] take advantage of the great success of AI chatbots, which are trained in visual question answering, image captioning, and other downstream tasks.

CLIP [30] demonstrates impressive visual representations trained together with paired text descriptions. The transformer encoder takes

the nonoverlapping patches and the words as input and utilizes the pooled encoded tokens to represent the images and sentences. 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 (image-text) encoder architecture is prevalent in multi-modal applications.

W. Kim *et al.* [18] utilizes the masked tokens in self-supervised learning in transformers [8] for natural language processing. The model takes tokenized sentences and image patches as input. Training tasks include paired classification and masked token filling. Kwon *et al.* [20] proposes the uniform transformer with two pre-training objectives, including masked vision and language modeling, and multi-modal alignment. Singh *et al.* [33] proposes the multi-modal encoder with visual and text encoders. The multi-modal encoder aligns the features from the two encoders with global contrastive learning and masked multi-modal modeling.

In addition to representation learning, the large language model provides incredible performance on text generation tasks. J. Yu *et al.* [46] optimize the visual encoder with image captioning as a downstream task. With a jointly trained visual encoder and language decoder, the model provides unified text and visual representations for the transformer. L. Yu *et al.* [47] employ the diffusion models [14, 34] for image generation and reinforce cross-modal representations. H. Liu *et al.* [25] combine the visual encoder with the LLM. The given image tokens are used as instructions for the detailed LLM responses. However, the training process requires large-scale paired images and texts, which is computationally expensive.

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 cannot understand the decision-making within fully connected layers because they lack semantic meanings. In addition to the lack of explainability, the fine-tuning process is computationally expensive. These models have billions of trainable parameters, and high-quality image-text pairs are required for tuning.

2.3 Green Learning

To handle the computationally intensive fine-tuning process and expand the image-text encoder using unpaired data, we introduce the Green Learning Alignment algorithm, which uses separately pre-trained image-text encoders. The idea of Green Learning was proposed by Kuo and Madni [19] and aims to reduce the computational cost of backpropagation while providing a theoretically explainable learning process for various applications. The modular designs can divide the problem into subproblems, which can be solved using transparent algorithms.

3 Proposed GEMMA Method

The GEMMA algorithm can be divided into three stages: 1) Global Alignment, 2) Image Cluster Alignment, and 3) Text Cluster Alignment. We adopt the multi-stage approach to approximate the complicated decision-making process rather than building a single large visual-language foundation model from scratch to ensure model efficiency. 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 single-layer adapter matrices to project the representations onto the joint latent space. Specifically, the alignment process consists of three modules: a) alignment, b) clustering in subdomains, and c) selection of subdomain features, where clustering and feature selection are performed in both the image and text domains, as shown in Fig. 2.

3.1 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 alignment module is illustrated in Fig. 3. The visual and text embeddings can be formulated as:

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

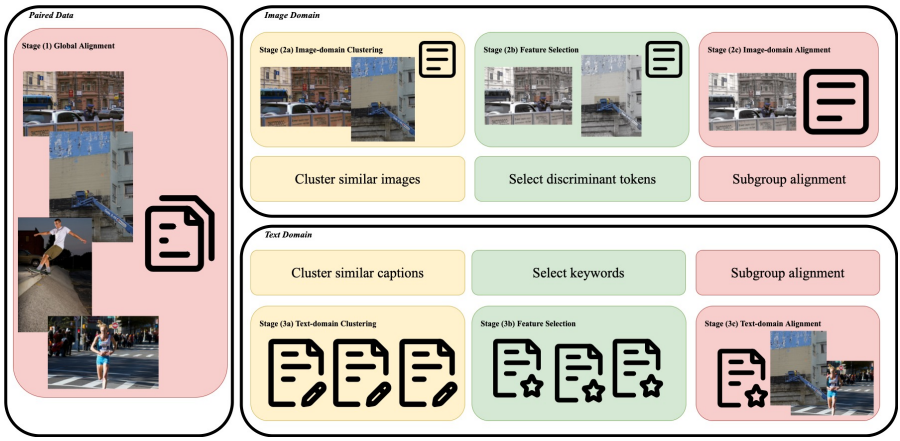

Figure 2: The overall algorithm design of Alignment. The first stage is the global alignment. The second and third stages include fine-grained clustering and feature selections in the image and text domain.

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 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 [4].

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

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 maximizes the similarity of relevant image-text pairs while avoiding negative image-text pairs from being embedded closely in the latent space.

Hence, the problem can be formulated as an optimization problem, and all 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.

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

Furthermore, we use the auxiliary matrices to constrain the joint representations and define the loss of cross-modality reconstruction as

$$\begin{aligned} \mathcal{L}_{cross-recon} = &\left\| C z_{txt} - e_{vis} \right\|_2 \\ &+ \left\| D z_{vis} - e_{txt} \right\|_2, \end{aligned} \tag{6}$$

where $C \in \mathbf{R}^{d_{vis} \times d_{joint}}$ and $C \in \mathbf{R}^{d_{txt} \times d_{joint}}$ are the auxiliary transformation matrices from the joint space onto the image and text modality, respectively. In addition, z_{vis}, z_{txt} are obtained from the corresponding paired caption or image data, e_{vis} and e_{txt}. However, the C and D matrices will not be used during inference. The alignment process is a linear transformation carried out by matrices A and B. The objective function can be written as:

$$\mathcal{L} = \alpha \mathcal{L}_{con} + \beta \mathcal{L}_{recon} + \gamma \mathcal{L}_{cross-recon}, \tag{7}$$

where α, β, and $\gamma \in \mathbf{R}$ represent hyperparameters in training. Linear alignment provides an 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 sub-datasets and utilize the stage-wise alignments for the detailed decision.

3.2 Sub-domain Clustering

With the alignment process, we can find similar representations by linear transformations. However, the transformation can only take global

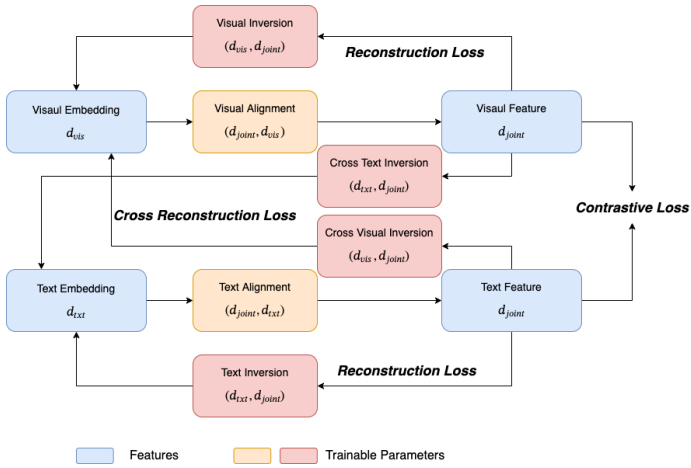

Figure 3: The illustration of the alignment process. The blue boxes are the features extracted by the frozen encoders. The orange boxes are the trainable transformation matrices. The red boxes are the auxiliary matrices for constraining the representations in the joint space.

representations, which means that images or captions are represented as d_{vis}- or d_{txt}-dimensional vectors. The image and sentence representations are the pooled output of the tokens in the prevailing transformer models. It can be inferred from previous research that fine-grained information is also crucial in information-matching tasks.

Due to the complexity of the fine-grained token representations, it is challenging to train the token-wise alignment in a brute-force manner. Thus, we adopt the clustering algorithms and use the clustering results to obtain crucial tokens. The crucial token selection will be introduced in Section 3.3. We can reduce the feature dimension from the number of tokens and perform a second-stage alignment.

We adopt frequency analysis and statistical approaches to construct a transparent and human-sensible intermediate structure. The clustering is conducted through (1) concept aggregation and (2) representation aggregation.

3.2.1 Concept Aggregation

We extract the concrete concepts for the candidate sentences by the Part-of-speech (POS) tagger [41]. We collect the nouns as anchors

and calculate the Term Frequency-Inverse Document Frequency (TF-IDF) to select the representative terms. As shown in Figure 4, the concepts lie in a long-tailed distribution, leading to a biased probability estimation. Hence, we aggregate the high-frequency terms based on the detector results and divide the candidate set into subsets for better-detailed alignments.

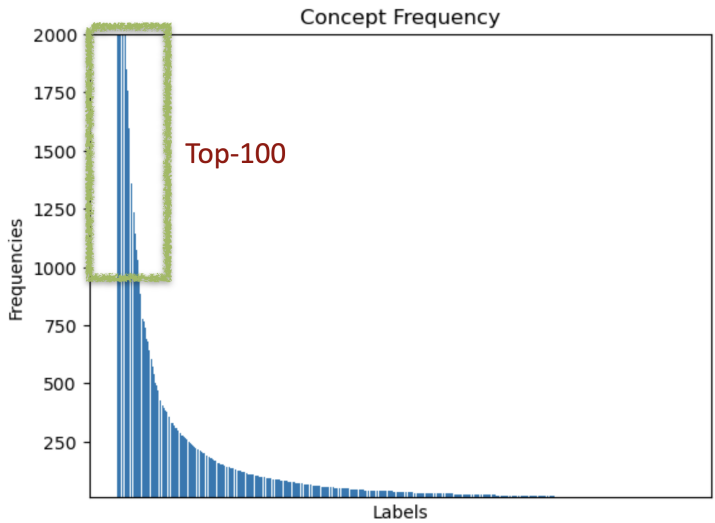

Figure 4: The frequency bar chart of the extracted corpus concepts. Top ten concepts and the corresponding counts are ('man', 36743), ('woman', 23845), ('people', 12810), ('shirt', 12743), ('girl', 10035), ('dog', 10030), ('boy', 9393), ('men', 8005), ('child', 7746), ('street', 7435), ('group', 6959), ('front', 6857), ('water', 5489), ('hat', 4075), ('person', 3810), ('ball', 3679), ('jacket', 3365), ('building', 3334), ('hand', 3113), and ('player', 3099).

We construct the co-occurrence matrix of the POS tagging and object detection results in the training set. As shown in Figure 5, concepts have a significant relationship with detection results. Hence, we can group the concepts tagged with POS with the probability conditional on the detection results. To visualize the physical meaning of the clusters, we can use the word clouds to show the high-frequency concepts in each cluster, shown in Figure 6.

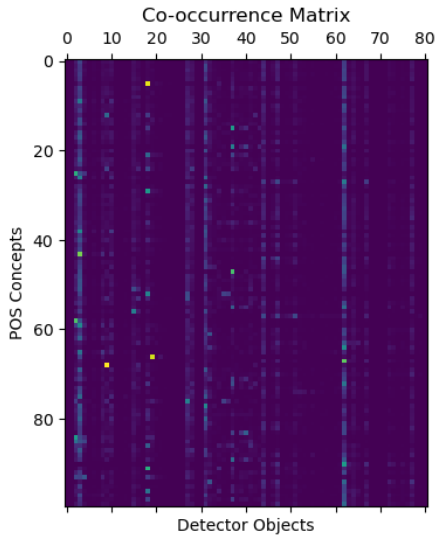

Figure 5: The occurrence matrix of POS tagging concepts and the detection results. The x-axis is the 80 object classes from the pretrained detector from the MS-COCO [23] dataset. The y-axis is the top 100 concepts from the POS tagger.

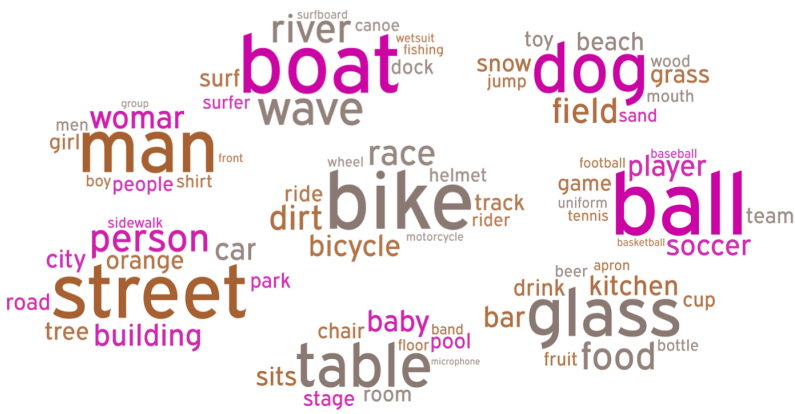

Figure 6: The visualization results of the clustering. The font size denotes the frequency of the word in the corpus.

3.2.2 Representation Aggregation

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

$$\|\tilde{u} - \tilde{v}\|_2^2 = \|\tilde{u}\|_2^2 + \|\tilde{v}\|_2^2 - 2\tilde{u}\tilde{v} = 2 - 2sim(u, v), \tag{8}$$

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

$$\begin{aligned} prob(u \in clus_i) &= \frac{e^{\epsilon'(2-2sim(u,cen_j))}}{\sum_{j=1}^{K} e^{\epsilon'(2-2sim(u,cen_i))}} \\ &= \frac{e^{\epsilon \cdot sim(u,cen_i)}}{\sum_{j=1}^{K} e^{\epsilon \cdot sim(u,cen_j)}}, \end{aligned} \tag{9}$$

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 ϵ increases, the probability distribution will concentrate on a certain class. If ϵ decreases, the probability distribution will become uniform.

We can group images and texts based on their probabilities and then align them using contrastive learning within these groups. We can improve the contrastive learning process by using negative samples similar to positive ones. We use hard-sample mining to ensure sample diversity within each group. The global alignment process helps identify the most challenging cases. We can then enlarge the groups by selecting the K-top candidates from the previous alignments as negative samples.

To clarify the roles of K-means clustering and the choice of hyperparameters, we conducted experiments comparing K-means and Agglomerative Clustering and varying the number of clusters. As shown in Table 1, increasing the number of clusters improves the retrieval in certain settings. However, this also requires training additional alignment matrices for the clusters. Therefore, we set the number of clusters to eight to strike a balance between the number of trainable parameters and the performance. K-means clustering is selected in GEMMA due to its slight empirical advantage over agglomerative clustering.

Table 1: Sensitivity to clustering methods, where R@k presents the top-k recalls and #Param denotes the number of trainable parameters. All the experiment is based on CLIP [30] visual encoder and RoBERTa [26] text encoder with Flickr30k [45] dataset.

Clustering	#Cluster	image-to-text			text-to-image			#Param
		R@1	R@5	R@10	R@1	R@5	R@10	
KMeans	4	84.1	95.7	96.6	65.3	90.1	93.4	5.2M
	8	86.3	**98.2**	99.4	73.2	**94.2**	97.2	10M
	16	**86.4**	98.1	**99.6**	**73.4**	94.2	**97.3**	20M
Agglomerative	4	84.0	94.4	96.2	64.8	90.0	92.2	5.2M
	8	85.5	97.7	98.7	72.9	92.8	96.1	10M
	16	86.0	96.9	99.5	73.4	93.7	97.0	20M

3.3 Feature Selection

Clustering results provide pseudo-labels for further feature selection. The label can be denoted as:

$$label^{u}_{clus_i} = \begin{cases} 0, & \text{if } prob(u \in clus_i) < T. \\ 1, & \text{otherwise.} \end{cases} \tag{10}$$

Here, $label^{u}_{clus_i}$ represents the label of the data point u whether it belongs to the group i, and $T \in (0, 1]$ is the self-definition threshold. With pseudo-labels, we can further adopt Discriminant Feature Selection [44] (DFT) to select informative features and reduce feature dimensions. DFT is a supervised feature selection process that measures dimension-wise importance. For a given 1D input feature, we can order the samples by the feature values and bind the feature dimension to the sample maximum and sample minimum. Then, we can partition the samples along the given dimension and calculate the partition purity by weighted cross-entropy with pseudo-labels obtained from Section 3.2. 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.

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, token-level alignments can be performed using the same procedure as global-level alignment.

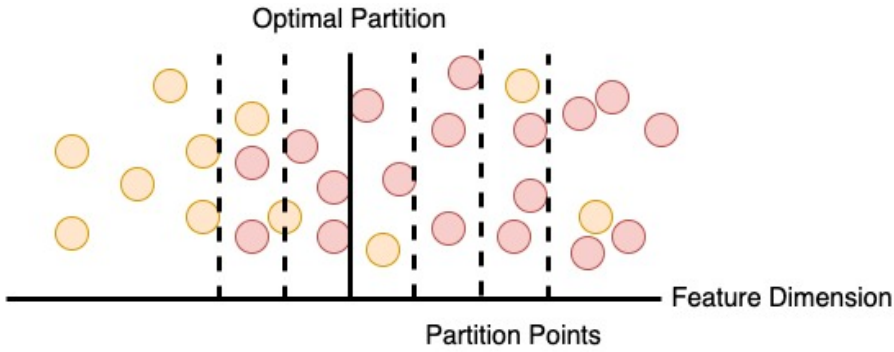

Figure 7: Visualization of DFT. Red and orange dots represent the binary labels. The partition metric is the weighted sum of the left and right binary cross-entropy. Dashed lines denote the potential partition points.

3.4 Mathematical Expression

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

$$\begin{aligned} & P(image|text) = P(image, text)/P(text) \\ & = P(text|image) \times \frac{P(image)}{P(text)} \\ & = \frac{1}{P(text)} \sum_{c \in cluster} P(text|image \in c) * P(image \in c) \\ & \propto \sum_{c \in cluster} P(text|image \in c) * P(image \in c), \end{aligned} \tag{11}$$

where $P(image|text)$ denotes the probability distribution of the images with a given query text. $P(image)$ and $P(text)$ denote the probability distribution of image and text, and *cluster* is the result of the clustering of our clustering modules. We further assume that the probability distribution within the cluster can be approximated as uniform. Conditional probability can reflect the stage-wise design in the proposed pipeline.

Furthermore, we use the similarity measurement to simplify the

probability estimator, which means that we use $sim(image, text)$ to represent $P(image|text)$. In the work, we adopt the cosine similarity as

$$\begin{aligned} & sim(image, text) \\ & = sim(W^{vis}\mathcal{F}(\text{image}), W^{txt}\mathcal{G}(\text{text})) \\ & = sim(W^{vis}[G_{vis}; T_{vis}], W^{txt}[G_{txt}; T_{txt}]) \\ & \sim sim(W^{vis}_{global}G_{vis}, W^{txt}_{global}G_{txt}) \\ & + sim(W^{vis}_{tokens}[G_{vis}; T_{vis}], W^{txt}_{global}G_{txt}) \\ & + sim(W^{vis}_{global}G_{vis}, W^{txt}_{tokens}[G_{txt}; T_{txt}]), \end{aligned} \tag{12}$$

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. The $[G; T]$ denotes the concatenated features of global and tokens. Due to 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 expressed as a combination of the conditional probabilities. For simplicity, we ignore the alignment matrix in the following representations.

$$\begin{aligned} & E[sim(\mathcal{F}(\text{image}), \mathcal{G}(\text{text}))] \\ & = E[E[sim(\mathcal{F}(\text{image}), \mathcal{G}(\text{text}))]|\text{image} \in C_1; \text{text} \in C_2] \\ & \sim E[sim(G_{vis}, G_{txt})] \\ & + E[E[sim(DFT([G_{vis}; T_{vis}]), G_{txt})|\text{image} \in C_1]] \\ & + E[E[sim(G_{vis}, DFT([G_{txt}; T_{txt}]))|\text{text} \in C_2]], \end{aligned} \tag{13}$$

where $DFT(.)$ represents the feature selection and dimension reduction process in Section 3.3 and C_1 and C_2 represent the cluster sets of K-means. Instead of training a complicated alignment process from the token-level output of the feature extractor, we propose a stage-wise 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 image and text domains.

4 Experiments

4.1 Dataset

We perform the image-to-text and text-to-image retrieval on the image-text benchmark: Flickr30k and MS-COCO. The Flickr30k dataset [45] 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 [23] is a larger-scale dataset with 123,287 images, each containing at least five captions. We follow the 'Karpathy' splitting for the experiments [17]: 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 refers to the top-K matches of the retrieval results. A retrieval is considered a true positive if the predicted matches include at least one of the paired ground-truth captions. Specifically, if the top K matches contain one of the five corresponding captions for a given image, it is counted as a positive in the recall metrics.

4.2 Hyperparameter Settings

The overall algorithm is trained stage by stage. We adopt K-means as the clustering algorithm. The number of clusters is 8, and ϵ is 50 for pseudo-labeling. For Flickr30K, we set the temperature parameter at 0.02, and the ratio between losses is set to $\alpha : \beta : \gamma = 1 : 0.5 : 0.6$ in the global alignment. In the alignment of the image subdomain, the temperature parameter is set to 0.015, and the ratio between the losses is set to $\alpha : \beta : \gamma = 1 : 0.4 : 0.5$. In the text subdomain alignment, the temperature parameter is set to 0.01, and the ratio between the losses is set to $\alpha : \beta : \gamma = 1 : 0.3 : 0.4$.

For the MS-COCO dataset, the temperature parameter is set to 0.05, and the ratio between the losses is set to $\alpha : \beta : \gamma = 1 : 0.5 : 0.5$ in the global alignment. In the alignment of the image subdomain, the temperature parameter is set to 0.03, and the ratio between the losses is set to $\alpha : \beta : \gamma = 1 : 0.3 : 0.5$. In the text subdomain alignment, the temperature parameter is set to 0.02, and the ratio between the losses is set to $\alpha : \beta : \gamma = 1 : 0.2 : 0.4$.

The dimension of the joint space is set to 768, which follows the token dimension of the transformer encoders. All optimization is performed using AdamW with the learning rate = 0.001.

4.3 Retrieval

Table 2: 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. [9] 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 [21]	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 [22]	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 [50]	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 [3]	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 [42]	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 [24]	76.4	94.3	97.3	57.4	82.3	89.0	–	–	–	–	–	–
SGRAF [10]	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 [16]	74.6	93.5	96.9	55.3	81.3	88.4	–	–	–	–	–	–
WCGL [40]	74.8	93.3	96.8	54.8	80.6	87.5	–	–	–	–	–	–
RCAR [9]	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 [15]	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 [30]	88.0	98.7	99.4	68.7	90.6	95.2	58.4	81.5	88.1	37.8	62.4	72.2
GEMMA(Ours)	**88.6**	**98.9**	**99.6**	**75.7**	**94.2**	**97.1**	58.6	83.2	90.0	**45.3**	**72.6**	**82.8**

We conducted the experiments and compared our alignment approach to the SOTA retrieval models. The results are shown in Table 2. We extract information from the frozen CLIP image and text encoder in the experiments. The CLIP encoders remain frozen during further alignments and serve as the baseline for our alignment process. 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 ($\sim 2.2\%$). The encoders remain untrainable during the training of alignment matrices. Therefore, GPU memory consumption is proportional to the trainable parameters, which can be reduced to less than 10 percent of the fully fine-tuned approach.

In Flickr30k (1k testing set), our approach outperforms other image-to-text and text-to-image retrieval methods. Alignment can improve recall @ 1 by 0. 6% in image-to-text retrieval. Meanwhile, our approach provides a 6% boost in text-to-image retrieval. RCAR [9] needs

dual-way optimized models, namely image-to-text and text-to-image. Our method is optimized in a feed-forward manner, and it ensembles the substructures directly.

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

4.4 Generalizability

This section demonstrates the alignment between the visual/text encoders, which are trained separately. The encoders remain frozen in the alignment process. All alignments are based on the grouping and linear projection proposed in our pipeline. The performance of CLIP visual and text encoders without GEMMA alignment is taken from the original CLIP paper. [30] Starting from the jointly trained CLIP structure, we change the text encoders into the RoBERTa [26] and the visual encoder into a CNN-based object detector [1]. All experiments are carried out on the Flickr30k dataset and follow the parameter settings in Section 4.2.

Table 3: 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 (GEMMA)	Flickr30k (1k testing set)					
			image-to-text			text-to-image		
			Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10
CLIP vis [30]	CLIP text [30]	x	88.0	98.7	99.4	68.7	90.6	95.2
DETR [2]	RoBERTa [26]	v	66.7	89.5	93.6	56.7	84.5	90.3
DETR [2]	CLIP text [30]	v	73.6	91.6	94.5	60.0	85.8	90.6
CLIP vis [30]	RoBERTa [26]	v	86.3	98.2	99.4	73.2	**94.2**	**97.2**
CLIP vis [30]	CLIP text [30]	v	**88.6**	**98.9**	**99.6**	**74.8**	**94.2**	97.1

The results are shown in Table 3. The best performance comes from the jointly trained models, whose representations are preliminarily aligned in the pre-training process. Compared to the CLIP visual encoder, the features of the object detector are weaker in the alignment process. However, the separately trained text encoder, RoBERTa [26], does not suffer from the unpaired training dataset. The representations from the CLIP visual encoder and the RoBERTa text encoder can provide competitive performance in image-to-text retrieval and 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 image and text data.

In contrast, the Convolution Neural Network (CNN)-based object detector representation cannot be applied directly to the image-text retrieval task. The decrease in performance results from global understanding. The object detector features are obtained from part of the image, and the representations lack a global understanding of the image. As to CLIP visual encoders, the visual tokens' pooled output contains the input images' global information and has detailed token features for us to process further stage alignments. The visual example can be found in Sec. 4.6. If the alignment process misses the global information in the very beginning, then the alignment process on detailed information may lead to a misfocused result.

4.5 Ablation Study on Different Stages

Due to the modularized design, we can compare the design from global alignment to subgroup alignment in the visual and textual domains. We choose encoders trained in different modalities to perform the alignment process. We use the CLIP visual encoder [30] and the RoBERTa [26] text encoder for the ablation study of stage-wise alignment on Flickr30k dataset [45]. The two encoders remain frozen in the experiments. The 'without alignment' setting means the direct dot product between the encoded features from two models. The two embeddings are located in different semantic latent spaces. Hence, the performance is the lowest compared to the other alignment processes.

With global alignment, the features can provide basic performance in retrieval tasks. However, a naive linear projection can not handle complex interactions between detailed information in the candidate set. Thus, recall rates increase as we add more stages in grouping, feature selection, and alignment. Feature selection provides statistical criteria for dimension reduction, preventing the latent dimension from increasing with additional tokens. We can take the essential features into the next stage and reduce computational cost simultaneously. Hence, the three-stage alignment can achieve the best performance with comparable efficiency.

Table 4: Ablation Studies on different stages, where R@k presents the top-k recalls and #Param denotes the number of trainable parameters. All the experiment is based on CLIP [30] visual encoder and RoBERTa [26] text encoder with Flickr30k [45] dataset.

Alignment	image-to-text			text-to-image		
	R@1	R@5	R@10	R@1	R@5	R@10
Without Alignment	64.5	71.7	84.3	32.7	61.6	80.1
Global	84.8	97.8	99.0	68.3	90.7	91.1
+Image Cluster	85.4	98.0	99.1	70.3	91.5	94.3
+Text Cluster (Final)	86.3	98.2	99.4	73.2	94.2	97.2

4.6 From Detection to Alignment

To better understand the difference between the visual features of transformers and object detectors, we demonstrate the retrieval processing step by step.

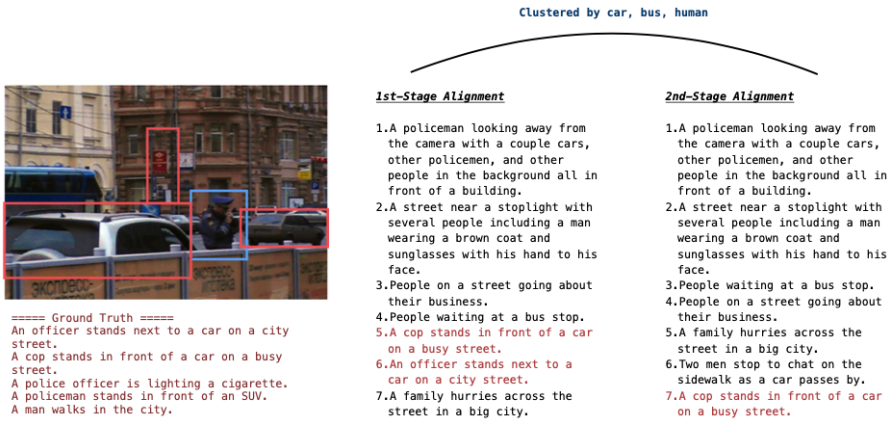

Figure 8: Error cases of object detector alignment. The object detector will give all objects equal weights and try to include all the objects in the captions.

The object detector can detect humans and vehicles, but the features lack a sensible relationship with each other. In the clustering stage, the clusters will focus on the specific object in the figure, that is, the bus in Fig. 8. In global alignment, the paired sentence is fifth. However, the correct captions fall to the seventh when we perform the finer alignment, which clusters on cars and buses. Although object de-

tectors can provide information fragments, the grouping process cannot link features. The detector features cannot find the central concept in the picture, but can be distracted by the surrounding objects.

The object detector can provide the features with the local information, yet the patched information is not represented in a structured manner. That is, we can only obtain the partial contents in the image and lose the global semantic representation in the clustering process. We rely on the global and local information relationship to retrieve suitable captions in the proposed coarse-to-fine clustering process.

On the other hand, the visual transformer can provide more information about the tokens and integrate the representations through a global pooling process. Hence, the token information can be selected in our feature selection module (sec. 3.3) and clustered according to the global features. The overall architecture can sort the rich representation in a coarse-to-fine manner and provide a better multi-modal alignment performance.

When comparing the alignment process across different features, it becomes evident that performance is influenced by the types of features used. However, the alignment process cannot transform weak visual features into strong ones. Instead, it aims to bridge the gap caused by differences in modality. Consequently, performance improves when encoded features have larger receptive fields. The proposed alignment does not require jointly fine-tuning the encoders in the limited paired multi-modal data and generalizes the single-modal encoder with additional alignment matrices.

Table 5: Experiments on Detector Features

Vis Feat		Text Feat	Flickr30k (1k testing set)					
			image-to-text			text-to-image		
Global Feat	Detail Feat		Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10
CLIP	CLIP	CLIP	85.3	91.9	93.3	72.1	90.6	92.2
DETR encoder	DETR decoder	CLIP	18.3	35.1	41.8	19.5	25.3	45.9
ResNet Backbone	DETR encoder	CLIP	66.7	89.5	93.3	56.7	84.5	90.3
ResNet Backbone	DETR decoder	CLIP	72.4	91.6	95.1	59.5	85.7	90.5
ResNet Backbone	DETR decoder	RoBERTa	64.5	84.5	88.4	53.3	83.3	87.3

5 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 subdomain matching and apply a feature selection method to decrease the input feature dimensions. All subprocesses can be expressed mathematically and analyzed statistically, providing transparency compared to black-box output. To ensure computational efficiency, we froze the visual and text encoders and only trained the alignment matrices, which represent only about 3% of the parameters compared to the original model.

In addition, we conducted experiments on applying our alignment mechanism to individually trained text and image encoders. In the testing dataset, we found that the pre-trained text encoder can improve the performance of text-to-image retrieval. Replacement of the text encoder can also lead to similar performance in image-to-text retrieval.

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

6 Acknowledgment

This work was supported by the DEVCOM Army Research Laboratory (ARL) under agreement W911NF2020157. Computation in the work was supported by the University of Southern California's Center for Advanced Research Computing (carc.usc.edu).

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang, "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, 6077–86.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, "End-to-end object detection with transformers", in *European conference on computer vision*, Springer, 2020, 213–29.

[3] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong 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, 12655–63.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations", in *International conference on machine learning*, PMLR, 2020, 1597–607.

[5] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, *et al.*, "Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks", *arXiv preprint arXiv:2312.14238*.

[6] Mengjun Cheng, Yipeng Sun, Longchao Wang, Xiongwei Zhu, Kun Yao, Jie Chen, Guoli Song, Junyu Han, Jingtuo Liu, Errui Ding, *et al.*, "Vista: Vision and scene text aggregation for cross-modal retrieval", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, 5184–93.

[7] 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.

[8] Jacob Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding", *arXiv preprint arXiv:1810.04805*.

[9] Haiwen Diao, Ying Zhang, Wei Liu, Xiang Ruan, and Huchuan Lu, "Plug-and-play regulators for image-text matching", *IEEE Transactions on Image Processing*.

[10] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu, "Similarity reasoning and filtration for image-text matching", in *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35, No. 2, 2021, 1218–26.

[11] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov, "Devise: A deep visual-semantic embedding model", *Advances in neural information processing systems*, 26.

[12] Tianyu Gao, Xingcheng Yao, and Danqi Chen, "Simcse: Simple contrastive learning of sentence embeddings", *arXiv preprint arXiv:2104.08821*.

[13] Raia Hadsell, Sumit Chopra, and Yann LeCun, "Dimensionality reduction by learning an invariant mapping", in *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, Vol. 2, IEEE, 2006, 1735–42.

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models", *Advances in neural information processing systems*, 33, 6840–51.

[15] Bhavin Jawade, Deen Dayal Mohan, Naji Mohamed Ali, Srirangaraj Setlur, and Venu 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, 1135–44.

[16] Zhong Ji, Kexin Chen, and Haoran Wang, "Step-wise hierarchical alignment network for image-text matching", *arXiv preprint arXiv:2106.06509*.

[17] Andrej Karpathy and Li Fei-Fei, "Deep visual-semantic alignments for generating image descriptions", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, 3128–37.

[18] Wonjae Kim, Bokyung Son, and Ildoo Kim, "Vilt: Vision-and-language transformer without convolution or region supervision", in *International conference on machine learning*, PMLR, 2021, 5583–94.

[19] C-C Jay Kuo and Azad M Madni, "Green learning: Introduction, examples and outlook", *Journal of Visual Communication and Image Representation*, 90, 103685.

[20] Gukyeong Kwon, Zhaowei Cai, Avinash Ravichandran, Erhan Bas, Rahul Bhotika, and Stefano Soatto, "Masked vision and language modeling for multi-modal representation learning", *arXiv preprint arXiv:2208.02131*.

[21] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He, "Stacked cross attention for image-text matching", in *Proceedings of the European conference on computer vision (ECCV)*, 2018, 201–16.

[22] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu, "Visual semantic reasoning for image-text matching", in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, 4654–62.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "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, 740–55.

[24] Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang, "Graph structured network for image-text matching", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, 10921–30.

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee, "Visual instruction tuning", *Advances in neural information processing systems*, 36.

[26] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized bert pretraining approach", *arXiv preprint arXiv:1907.11692*.

[27] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim, "Dual attention networks for multimodal reasoning and matching", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, 299–307.

[28] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese, "Deep metric learning via lifted structured feature embedding", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, 4004–12.

[29] Jeffrey Pennington, Richard Socher, and Christopher D Manning, "Glove: Global vectors for word representation", in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, 1532–43.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, *et al.*, "Learning transferable visual models from natural language supervision", in *International conference on machine learning*, PMLR, 2021, 8748–63.

[31] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka, "Contrastive learning with hard negative samples", *arXiv preprint arXiv:2010.04592*.

[32] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, 815–23.

[33] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela, "Flava: A foundational language and vision alignment model", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, 15638–50.

[34] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics", in *International conference on machine learning*, PMLR, 2015, 2256–65.

[35] Kihyuk Sohn, "Improved deep metric learning with multi-class n-pair loss objective", *Advances in neural information processing systems*, 29.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need", *Advances in neural information processing systems*, 30.

[37] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik, "Learning two-branch neural networks for image-text matching tasks", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), 394–407.

[38] Peng Wang, Qi Wu, Jiewei Cao, Chunhua Shen, Lianli Gao, and Anton van den Hengel, "Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, 1960–8.

[39] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, *et al.*, "Image as a foreign language: Beit pretraining for all vision and vision-language tasks", *arXiv preprint arXiv:2208.10442*.

[40] Yun Wang, Tong Zhang, Xueya Zhang, Zhen Cui, Yuge Huang, Pengcheng Shen, Shaoxin Li, and Jian Yang, "Wasserstein coupled graph learning for cross-modal retrieval", in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, 2021, 1793–802.

[41] Chengwei Wei, Runqi Pang, and C-C Jay Kuo, "GWPT: A Green Word-Embedding-based POS Tagger", *arXiv preprint arXiv:2401.07475*.

[42] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng 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, 10941–50.

[43] Hong Xuan, Abby Stylianou, Xiaotong Liu, and Robert Pless, "Hard negative examples are hard, but useful", in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, Springer, 2020, 126–42.

[44] Yijing Yang, Wei Wang, Hongyu Fu, C-C Jay Kuo, *et al.*, "On supervised feature selection from high dimensional feature spaces", *APSIPA Transactions on Signal and Information Processing*, 11(1).

[45] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions", *Transactions of the Association for Computational Linguistics*, 2, 67–78.

[46] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu, "Coca: Contrastive captioners are image-text foundation models", *arXiv preprint arXiv:2205.01917*.

[47] Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, *et al.*, "Scaling autoregressive multi-modal models: Pretraining and instruction tuning", *arXiv preprint arXiv:2309.02591*, 2(3).

[48] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi, "From recognition to cognition: Visual commonsense reasoning", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, 6720–31.

[49] Yan Zeng, Xinsong Zhang, Hang Li, Jiawei Wang, Jipeng Zhang, and Wangchunshu Zhou, "X 2-vlm: All-in-one pre-trained model for vision-language tasks", *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[50] Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z Li, "Context-aware attention network for image-text retrieval", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, 3536–45.

[51] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen, "Dual-path convolutional image-text embeddings with instance loss", *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(2), 1–23.

[52] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler, "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, 19–27.