DALR: Dual-level Alignment Learning for Multimodal Sentence Representation Learning

Kang He, Yuzhe Ding, Haining Wang, Fei Li, Chong Teng*, Donghong Ji
Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University
{hekang0225,lifei_csnlp,tengchong}@whu.edu.cn

Abstract

Previous multimodal sentence representation learning methods have achieved impressive performance. However, most approaches focus on aligning images and text at a coarse level, facing two critical challenges: *cross-modal misalignment bias* and *intra-modal semantic divergence*, which significantly degrade sentence representation quality. To address these challenges, we propose **DALR** (Dual-level Alignment Learning for Multimodal Sentence Representation). For cross-modal alignment, we propose a consistency learning module that softens negative samples and utilizes semantic similarity from an auxiliary task to achieve fine-grained cross-modal alignment. Additionally, we contend that sentence relationships go beyond binary positive-negative labels, exhibiting a more intricate ranking structure. To better capture these relationships and enhance representation quality, we integrate ranking distillation with global intra-modal alignment learning. Comprehensive experiments on semantic textual similarity (STS) and transfer (TR) tasks validate the effectiveness of our approach, consistently demonstrating its superiority over state-of-the-art baselines.

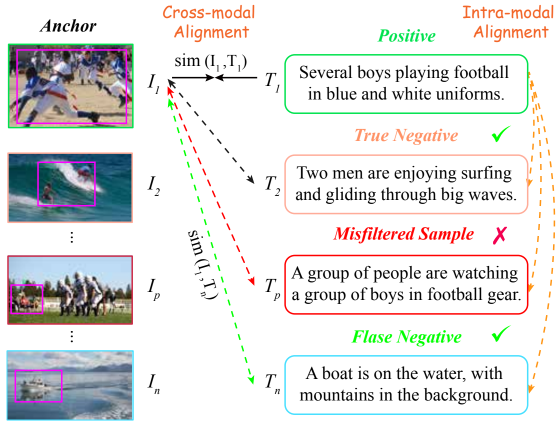

Figure 1: Illustration of a batch image-caption pairs from the Flickr dataset. KDMCSE sets a threshold based on $sim(I, T)$ to filter out false negatives. ✓: denotes the sample is correctly classified as false negative or true negative based on image-text similarity. ✗: indicates a sample misclassified as a false negative and erroneously filtered due to its high similarity with the anchor image.

1 Introduction

Sentence representation learning converts sentences into low dimensional vectors to preserve semantic information and is widely used in NLP tasks, such as semantic similarity (Agirre et al., 2012, 2013), information extraction (Wang et al., 2022a; Zheng et al., 2024), and content analysis (Ling et al., 2022; Wang et al., 2024; Zheng et al., 2025). With the success of pre-trained language models (PLMs) such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), numerous methods (Gao et al., 2021; Wu et al., 2022b; Zhang et al., 2022b; He et al., 2023; Seonwoo et al., 2023; He et al., 2025) have achieved remarkable performance by contrastive learning and different augmentation strategies.

Unfortunately, the methods of constructing positive (Yan et al., 2021; Wu et al., 2022a; Zhuo et al., 2023) and negative (Zhou et al., 2022; Deng et al., 2023; Shi et al., 2023) samples are usually too simple to capture nuanced semantic relationships between sentences deeply. For example, although "*A man is skating*" and "*A man is gliding*" are mutually exclusive in common sense, this contradiction is not easily captured through text alone. However, visual information can naturally reveal such contradictions, providing a rich supervision signal for better understanding (Wang et al., 2022b). Incorporating visual signals into language models has been shown to improve performance across various downstream tasks (Bordes et al., 2020; Tang et al., 2021; Nguyen et al., 2023; Huang et al., 2023a). MCSE (Zhang et al., 2022a) leveraged multimodal contrastive learning for cross-modal alignment, and

* Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2025, pages 3586–3601
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

KDMCSE (Nguyen et al., 2024) further enhanced alignment by filtering highly similar samples to reduce false negatives and applying adaptive angular contrastive learning to better distinguish negatives. Despite these advances, aligning text and image through semantic similarity still faces two key challenges: *cross-modal misalignment bias* and *intra-modal semantic divergence*.

Cross-modal Misalignment Bias (CMB) stems from the inherent asymmetry between modalities when aligning image-text pairs. Text is typically information-dense and selective, focusing on key details, while images capture all components indiscriminately, leading to significant redundancy. As shown in the purple box in Figure 1, the focal object of the image I_n is "a boat", which occupies only a small region, with most visual patches containing irrelevant information. Moreover, due to cognitive biases among annotators, a single image may have multiple semantic descriptions (Chun et al., 2021), further amplifying the heterogeneity between modalities. This mismatch causes semantic similarity to misrepresent true alignment, resulting in biased representations.

Intra-modal Semantic Divergence (ISD) refers to the erroneous identification of semantically divergent texts as highly similar due to their shared reference to the same image. Studies (Chun et al., 2022; Parekh et al., 2021) have noted that multiple captions (or images) can describe the same image (or caption) with differing focuses. For instance, in Figure 1, given the anchor image I_1, caption T_1 ("*Several boys playing football in blue and white uniforms*") emphasizes the players' appearance and activity, while T_p ("*A group of people are watching a group of boys in football gear*") highlights the spectators. Despite their semantic divergence, both captions exhibit high image similarity, resulting in false negatives. This misalignment undermines intra-modal consistency and degrades sentence representation quality.

To address these challenges, we propose **DALR**: a **D**ual-level **A**lignment **L**earning Framework for Multimodal Sentence **R**epresentation. First, for cross-modal alignment, we introduce an auxiliary cross-modal consistency task that enhances supervision by predicting image-text correspondence through a binary classification framework. This task extracts latent semantic features and constructs a semantic similarity matrix as a soft target to unify representations across modalities. Second, to mitigate intra-modal semantic divergence, we argue that sample relationships are inherently continuous rather than binary. We propose an intra-modal alignment strategy, employing multi-teacher models to generate coarse-grained semantic rankings as pseudo-labels. This strategy incorporates KL divergence to ensure the student model captures global information from the teachers, thereby achieving robust intra-modal alignment.

Experiments on the widely-used STS and TR tasks showcase the considerable effectiveness of DALR. Ablation studies and visualization analysis further validate the existence of CMB and ISD issue and the necessity of joint modality alignment. The main contributions are summarized as follows:

- We introduce DALR to enhance text representations through joint cross-modal and intra-modal alignment.
- We propose a cross-modal alignment method with auxiliary tasks to soften negative samples and improve alignment to mitigate CMB issue.
- We adopt ranking distillation with global alignment learning to capture fine-grained semantic structures for ISD issue.
- Thorough experiments show that DALR improves the performance over all metrics and achieves state-of-the-art on two benchmarks[1].

2 Related Work

2.1 Sentence Representation Learning

Sentence representation learning is a fundamental task in natural language processing. Early methods, such as Skip-Thought (Kiros et al., 2015) and FastSent (Hill et al., 2016), leverage contextual relationships to learn sentence representations. With the progression of PLMs and SimCSE (Gao et al., 2021), the "PLMs + contrastive learning" paradigm has become increasingly prevalent. Data augmentation strategies (Yan et al., 2021; Wu et al., 2022b; Zhuo et al., 2023; He et al., 2023) enhance representation quality by generating diverse positive samples. ConSERT (Yan et al., 2021) uses dropout masking and token shuffling, while PCL (Wu et al., 2022a) adopts multiple augmentation techniques. WhitenedCSE (Zhuo et al., 2023) improves diversity through inter-group whitening. Additionally, advancements in negative sampling (Zhou et al.,

[1] https://github.com/Hekang001/DALR.

2022; Deng et al., 2023) and hard negative construction (Shi et al., 2023) further refine sentence representation learning.

2.2 Modality Alignment

Research on modality alignment (Cheng et al., 2023b; Liu et al., 2023b; Zhang et al., 2023; Han et al., 2024) aims to unify feature representations across modalities (e.g., image, text, audio) for enhanced representation learning (Li et al., 2021; Huang et al., 2023b; Zhu et al., 2023), cross-modal understanding (Yu et al., 2023; Li et al., 2023), and generation tasks (Sung-Bin et al., 2023; Tian et al., 2023). Methods like ALBEF (Li et al., 2021) align image-text features through cross-modal attention, while MVPTR (Li et al., 2022) focuses on multi-level semantic alignment. MCSE (Zhang et al., 2022a) integrates visual information into sentence embeddings, and KDMCSE (Nguyen et al., 2024) improves this by leveraging external models for distillation and filtering false negatives. In contrast, our approach balances cross-modal alignment with intra-modal semantic consistency, enhancing visual information utilization and improving sentence representation quality.

3 Methodology

3.1 Preliminary Work

Unsupervised SimCSE Unsupervised SimCSE (Gao et al., 2021) leverages dropout as a minimal data augmentation strategy. Given a sentence set $T = \{t_i\}_{i=1}^{m}$, each sentence is encoded twice with different dropout masks, producing two representations $s_i^z = g_{\varphi_\theta}(f_\theta(t_i, z))$ and $s_i^{z'} = g_{\varphi_\theta}(f_\theta(t_i, z'))$, where f_θ is a pre-trained language encoder (e.g., BERT), and g_{φ_θ} is a projection head. The [CLS] token is used as the final embedding, and the objective is to maximize the similarity between paired representations:

$$\mathcal{L}_{text} = -\sum_{i=1}^{N} \log \frac{e^{sim\left(s_i^z, s_i^{z'}\right)/\tau}}{\sum_{j=1}^{N} e^{sim\left(s_i^z, s_j^{z'}\right)/\tau}} \tag{1}$$

where N is the batch size and τ is a temperature hyper-parameter. $sim(\boldsymbol{h}_1, \boldsymbol{h}_2) = \frac{\boldsymbol{h}_1^T \boldsymbol{h}_2}{\|\boldsymbol{h}_1\| \cdot \|\boldsymbol{h}_2\|}$ is cosine similarity function.

Multimodal Contrastive Learning Given a set of image-text pairs represented as $C = \{v_i, t_i\}_{i=1}^{N} \in \mathcal{D}$, MCSE (Zhang et al., 2022a) projects text t_i and image v_i into a unified space:

$$s_i^z = g_{\varphi_\theta}(f_\theta(t_i, z)) \tag{2}$$

$$h_i^v = g_{\varphi_v}(f_v(v_i)), \quad h_i^t = g_{\varphi_t}(f_t(t_i)) \tag{3}$$

where $f_v(\cdot)$ denotes a frozen image teacher encoder, and $f_t(\cdot)$ refers to a frozen text teacher encoder. (More details for image and text teacher encoder are in Section 4.1 and Appendix B.) z denotes the dropout mask, $g_{\varphi_\theta}(\cdot)$ is the projection head of the language student model that projects the sentence representation into a shared space, $g_{\varphi_v}(\cdot)$ and $g_{\varphi_t}(\cdot)$ are the projection heads of the image and text teacher models, respectively. Therefore, the multimodal contrastive learning objective using InfoNCE (Oord et al., 2018) is expressed as:

$$\mathcal{L}_{Info} = -\sum_{i=1}^{N} \log \frac{e^{sim\left(s_i^z, h_i^v\right)/\tau}}{\sum_{j=1}^{N} e^{sim\left(s_i^z, h_j^v\right)/\tau}} \tag{4}$$

3.2 Cross-modal Alignment learning

Figure 2 illustrates the main workflow of DALR. Image and text features exhibit a significant semantic gap, making direct mapping into a shared space for alignment challenging. We propose a cross-modal alignment method with an auxiliary consistency task to capture fine-grained image-text semantics. The generated similarity matrix refines negative samples, providing a guiding signal for enhanced cross-modal contrastive learning.

Cross-modal consistency learning We formulate this module as a binary classification task to predict image-text alignment based on multimodal features. Given the original dataset $\mathcal{D}$ with aligned image-text pairs, we construct a new dataset $\mathcal{D}'$ by shuffling images to create mismatched pairs. This enables the model to learn to distinguish between aligned and misaligned pairs. For each image-text pair $C' = \{v', t'\} \in \mathcal{D}'$, we extract unimodal representations using f_v and f_θ, which are then projected into a shared space via modality-specific MLPs, obtaining shared representations $h_s^{v'}$ and $s_s^{z'}$ as defined in Eq.2 and Eq.3. We use the cosine embedding loss function with margin m for optimization as follows:

$$\mathcal{L}_{cons} = \begin{cases} 1 - \cos(h_s^{v'}, s_s^{z'}) & \text{if } y' = 1, \\ \max(0, \cos(h_s^{v'}, s_s^{z'}) - m) & \text{if } y' = 0. \end{cases} \tag{5}$$

where $\cos(\cdot)$ represents the normalized cosine similarity, and m controls the margin for negative samples, typically set to 0.2 based on empirical findings. The consistency learning task captures deeper semantic relationships by refining the matching between images and texts. It enhances the model's

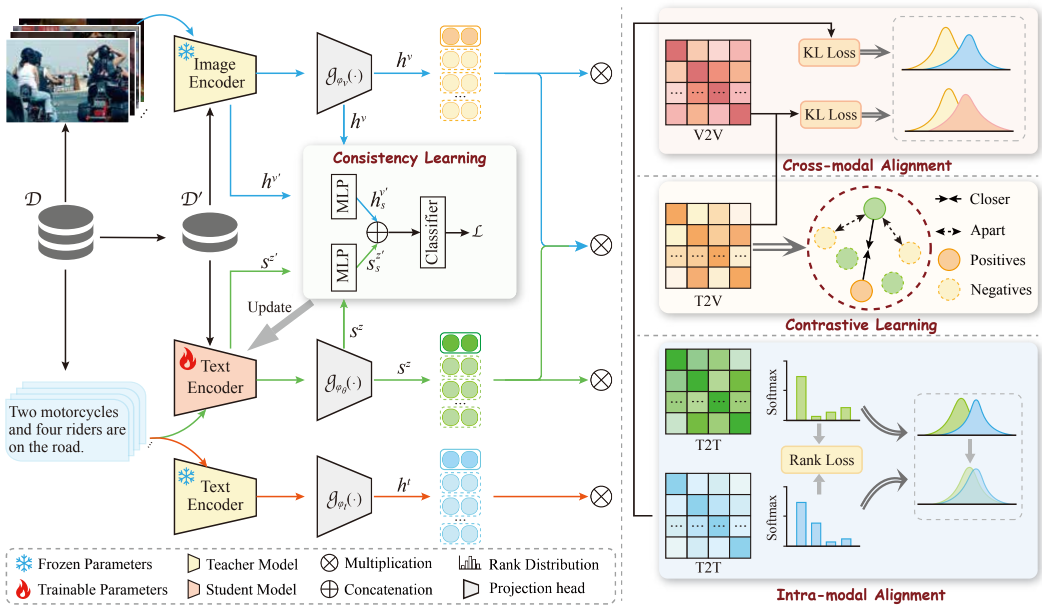

Figure 2: The illustration of our proposed framework DALR, consisting of three components: (a) the multimodal constrastive learning module uses the guidance of visual information to obtain sentence representations, (b) the cross-modal alignment module further aligns cross-modal features, and (c) the intra-modal alignment module enhances internal alignment through ranking distillation learning and KL divergence.

multimodal representation, improves the discrimination of negative samples, and reduces noise. Notably, this task can be learned in parallel with contrastive learning, generating cross-modal soft labels to guide alignment.

Cross-modal alignment We use the representation s_i^z obtained by the language student model and h_i^v obtained by the visual teacher model to calculate the cosine similarity, and perform normalization to obtain the probability distribution P_{ij}^{v2t} of pairing v_i with t_j:

$$P_{ij}^{t2v} = \frac{e^{sim(s_j^z, h_i^v)}}{\sum_{k=1}^{N} e^{sim(s_k^z, h_i^v)}}, P_i^{t2v} = \left(P_{i1}^{t2v}, P_{i2}^{t2v}, ..., P_{iN}^{t2v}\right) \quad (6)$$

where P_i^{t2v} is the probability distribution set composed of P_{ij}^{t2v} in the same batch. At the same time, we compute the cosine similarity within the teacher text model and normalize it to obtain the probability estimate Q_{ij}^{t2t} from the teacher model:

$$Q_{ij}^{t2t} = \frac{e^{sim(h_i^t, h_j^t)}}{\sum_{j=1}^{N} e^{sim(h_i^t, h_j^t)}}, Q_i^{t2t} = \left(Q_{i1}^{t2t}, Q_{i2}^{t2t}, ..., Q_{iN}^{t2t}\right) \quad (7)$$

Similarly, the similarity between h_i^v and h_j^v is calculated through the features obtained by the teacher vision model to obtain Q_i^{v2v}. In model training, we promote the alignment between images and texts by minimizing the KL divergence between the target distribution (Q_i^{t2t}, Q_i^{v2v}) and the predicted distribution P_i^{t2v}:

$$\mathcal{L}_{CMA} = \frac{1}{2}\sum_{i=1}^{N}(D_{KL}(Q_i^{t2t}||P_i^{t2v}) + D_{KL}\left(Q_i^{v2v}||(P_i^{t2v})^T\right) \quad (8)$$

where $D_{KL}(\cdot)$ represents KL divergence. Minimizing KL divergence is equivalent to maximizing the mutual information between the teacher and student distributions, which facilitates cross-modal information transfer to some extent. Through the aforementioned two parts, we can capture more cross-modal detailed semantic information and facilitate the learning of sentence representation. The final loss $\mathcal{L}_{CML}$ of cross-modal contrastive learning is calculated as follows:

$$\mathcal{L}_{CML} = \mathcal{L}_{cons} + \mathcal{L}_{CMA} \quad (9)$$

3.3 Intra-modal Alignment Learning

Despite progress in cross-modal alignment, existing methods, such as KDMCSE (Nguyen et al., 2024), overlook the sparsity of image information and the variation in textual focus on different image regions. This results in multiple low-similarity texts aligning with a single image (Chun et al., 2021; Parekh et al., 2021), undermining semantic accuracy, a phenomenon we term "*intra-modal semantic divergence*".

To address this, we introduce an intra-modal alignment method featuring two components: ranking distillation for fine-grained semantic capture and KL-based inter-modal alignment for global distribution learning. Ranking information, which reflects subtle structural differences between sentences, enhances modality alignment. We employ multiple teachers (SimCSE and DiffCSE) to provide comprehensive ranking data, with a weighted combination of [CLS] token embeddings yielding the final representation. The teachers' similarity score lists act as pseudo-ranking labels, guiding the intra-modal alignment. We apply ListMLE (Xia et al., 2008) to refine ranking learning:

$$\mathcal{L}_{rank} = -\sum_{i=1}^{N} \log \left(\prod_{j=1}^{M} \frac{\exp\left(S(x_i)_{\pi_i^T(j)}/\tau\right)}{\sum_{k=j}^{M} \exp\left(S(x_i)_{\pi_i^T(k)}/\tau\right)} \right) \tag{10}$$

where $S(x_i)$ represents the list of similarity scores generated by the student model for the text input x_i, $\pi_i^T(j)$ is the index of the j-th position in the ranking π_i^T generated by the teacher model, and $S(x_i)_{\pi_i^T(j)}$ represents the score of the student model for the j-th position in the ranking.

ListMLE directly optimizes ranking order but neglects the probabilistic structure of the score distribution. This simplified approach may fail to capture the global probability information from the teacher model, limiting sentence representation performance. To address this, we introduce KL divergence to minimize the statistical distribution gap between the teacher and student models, aligning pseudo-labels with the student model's predictions. This reduces confusion between pseudo-labels and model outputs, enhancing learning effectiveness. Specifically, using Eq.6, we can derive the text distribution probability P_i^{t2t} of the student model:

$$P_{ij}^{t2t} = \frac{e^{sim(s_i^z, s_i^{z'})}}{\sum_{j=1}^{N} e^{sim(s_i^z, s_j^{z'})}}, P_i^{t2t} = (P_{i1}^{t2t}, P_{i2}^{t2t}, ..., P_{iN}^{t2t}) \tag{11}$$

where z, z' represent different dropouts, and P_i^{t2t} is a probability distribution set consisting of a set of probability distributions $\mathcal{P} = \{P_{ij}^{t2t}\}_{j=1}^{N}$. Finally, we learn a more general distribution by optimizing the KL divergence between the teacher distribution probability Q_i^{t2t} and the student distribution probability P_i^{t2t}. The objective is as follows:

$$\mathcal{L}_{IMA} = \sum_{i=1}^{N} (D_{KL}(Q_i^{t2t} || P_i^{t2t}) \tag{12}$$

By combining $\mathcal{L}_{rank}$ and $\mathcal{L}_{IMA}$, we can ensure that the student model not only matches the overall similarity distribution (KL divergence), but also preserves the critical ranking information (ListMLE). Therefore, the goal of intra-modal alignment learning is:

$$\mathcal{L}_{IML} = \mathcal{L}_{rank} + \mathcal{L}_{IMA} \tag{13}$$

3.4 Training Objectives

According to Eq.4, Eq.9 and Eq.13, we can add all losses to a final loss:

$$\mathcal{L}_{total} = \mathcal{L}_{Info} + \lambda \mathcal{L}_{CML} + \mu \mathcal{L}_{IML} \tag{14}$$

where λ and μ are hyper-parameters for weights balance.

4 Experiments

4.1 Experiments Setup

We evaluate our method on two sentence related tasks: semantic textual similarity (STS) and transfer (TR) task. For the STS tasks, we evaluate on seven datasets: STS 2012-2016 (Agirre et al., 2012, 2013, 2014, 2015, 2016), STS Benchmark (Cer et al., 2017) and SICK-Relatedness (Marelli et al., 2014). We use the SentEval toolkit (Conneau and Kiela, 2018) for evaluation and adopt the Spearman's correlation coefficient (multiplied by 100) as the reporting metric. For the TR tasks, we also use SentEval to evaluate on seven datasets: MR (Pang and Lee, 2005), CR (Hu and Liu, 2004), SUBJ (Pang and Lee, 2004), MPQA (Wiebe et al., 2005), SST-2 (Socher et al., 2013), TREC (Voorhees and Tice, 2000) and MRPC (Dolan and Brockett, 2005).

Datasets According to MCSE (Zhang et al., 2022a), we use Flickr (Young et al., 2014) and MSCOCO (Lin et al., 2014) as multimodal sentence embedding datasets. In addition, we follow SimCSE (Gao et al., 2021) and use 1,000,000 sentences randomly selected from Wikipedia as the training dataset.

Baseline Models Following the standard protocol on the two benchmarks (Gao et al., 2021), we compare our model with three baseline models: SimCSE (Gao et al., 2021), MSE (Zhang et al., 2022a), KDMCSE (Nguyen et al., 2024). More details of baseline models are in Appendix A.

Implementation Details During model initialization, we utilize SimCSE and DiffCSE as two text teachers and load the checkpoint of CLIP-ViT-B/32 as the image teacher model. During training, considering that the sizes of the pure-text dataset (with

	Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.↑
wiki	SimCSE-BERT♡	67.8$_{\pm 1.6}$	80.0$_{\pm 2.1}$	72.5$_{\pm 1.7}$	80.1$_{\pm 0.8}$	77.6$_{\pm 0.8}$	76.5$_{\pm 0.8}$	70.1$_{\pm 0.9}$	74.9$_{\pm 1.1}$
	SimCSE-RoBERTa♡	68.7$_{\pm 1.0}$	82.0$_{\pm 0.5}$	74.0$_{\pm 1.0}$	82.1$_{\pm 0.4}$	81.1$_{\pm 0.4}$	80.6$_{\pm 0.3}$	69.2$_{\pm 0.2}$	76.8$_{\pm 0.5}$
wiki+flickr	SimCSE-BERT†	69.9$_{\pm 1.7}$	79.8$_{\pm 1.5}$	72.9$_{\pm 0.9}$	81.9$_{\pm 0.8}$	77.8$_{\pm 0.9}$	76.6$_{\pm 1.1}$	68.4$_{\pm 0.8}$	75.3$_{\pm 0.9}$
	MCSE-BERT†	71.4$_{\pm 0.9}$	81.8$_{\pm 1.3}$	74.8$_{\pm 0.9}$	83.6$_{\pm 0.9}$	77.5$_{\pm 0.8}$	79.5$_{\pm 0.5}$	72.6$_{\pm 1.4}$	77.3$_{\pm 0.5}$
	KDMCSE-BERT‡	**74.4**$_{\pm 1.4}$	83.1$_{\pm 0.9}$	76.3$_{\pm 1.1}$	83.7$_{\pm 0.8}$	78.8$_{\pm 0.9}$	81.3$_{\pm 0.9}$	73.0$_{\pm 0.9}$	78.6$_{\pm 0.8}$
	DALR-BERT	73.9$_{\pm 0.8}$	**84.0**$_{\pm 0.7}$	**76.5**$_{\pm 0.5}$	**84.3**$_{\pm 0.9}$	**80.6**$_{\pm 1.1}$	**81.8**$_{\pm 0.2}$	**75.3**$_{\pm 0.4}$	**79.5**$_{\pm 0.7}$
	SimCSE-RoBERTa†	69.5$_{\pm 0.9}$	81.6$_{\pm 0.5}$	74.1$_{\pm 0.6}$	82.4$_{\pm 0.3}$	80.9$_{\pm 0.5}$	79.9$_{\pm 0.3}$	67.3$_{\pm 0.5}$	76.5$_{\pm 0.4}$
	MCSE-RoBERTa†	71.7$_{\pm 0.2}$	82.7$_{\pm 0.4}$	75.9$_{\pm 0.3}$	84.0$_{\pm 0.4}$	81.3$_{\pm 0.3}$	82.3$_{\pm 0.5}$	70.3$_{\pm 1.3}$	78.3$_{\pm 0.1}$
	KDMCSE-RoBERTa‡	**73.6**$_{\pm 0.7}$	83.8$_{\pm 0.6}$	**77.4**$_{\pm 0.4}$	84.0$_{\pm 0.3}$	81.5$_{\pm 0.7}$	82.3$_{\pm 0.6}$	71.2$_{\pm 0.4}$	79.1$_{\pm 0.3}$
	DALR-RoBERTa	**73.6**$_{\pm 0.4}$	**84.4**$_{\pm 0.2}$	77.2$_{\pm 0.6}$	**84.9**$_{\pm 0.7}$	**82.0**$_{\pm 0.4}$	**82.6**$_{\pm 0.2}$	**74.6**$_{\pm 0.7}$	**79.9**$_{\pm 0.5}$
wiki+coco	SimCSE-BERT†	69.1$_{\pm 1.0}$	80.4$_{\pm 0.9}$	72.7$_{\pm 0.7}$	81.1$_{\pm 0.3}$	78.2$_{\pm 0.9}$	73.9$_{\pm 0.6}$	66.6$_{\pm 1.2}$	74.6$_{\pm 0.2}$
	MCSE-BERT†	71.2$_{\pm 1.3}$	79.7$_{\pm 0.9}$	73.8$_{\pm 0.9}$	83.0$_{\pm 0.4}$	77.8$_{\pm 0.9}$	78.5$_{\pm 0.4}$	72.1$_{\pm 1.4}$	76.6$_{\pm 0.5}$
	KDMCSE-BERT‡	73.2$_{\pm 1.2}$	80.5$_{\pm 1.0}$	75.4$_{\pm 0.9}$	83.2$_{\pm 0.3}$	79.7$_{\pm 0.8}$	79.7$_{\pm 0.7}$	73.7$_{\pm 1.4}$	77.9$_{\pm 1.2}$
	DALR-BERT	**73.4**$_{\pm 1.0}$	**82.6**$_{\pm 1.2}$	**75.6**$_{\pm 0.8}$	**83.5**$_{\pm 0.6}$	**80.8**$_{\pm 0.7}$	**80.5**$_{\pm 0.5}$	**74.1**$_{\pm 0.9}$	**78.6**$_{\pm 0.9}$
	SimCSE-RoBERTa†	66.4$_{\pm 0.9}$	80.7$_{\pm 0.7}$	72.7$_{\pm 1.1}$	81.3$_{\pm 0.9}$	80.2$_{\pm 0.8}$	76.8$_{\pm 0.6}$	65.7$_{\pm 0.7}$	74.8$_{\pm 0.5}$
	MCSE-RoBERTa†	70.2$_{\pm 1.7}$	82.0$_{\pm 0.7}$	75.5$_{\pm 1.2}$	83.0$_{\pm 0.6}$	81.5$_{\pm 0.7}$	80.8$_{\pm 1.0}$	69.9$_{\pm 0.6}$	77.6$_{\pm 0.8}$
	KDMCSE-RoBERTa‡	72.8$_{\pm 1.5}$	81.7$_{\pm 0.9}$	76.1$_{\pm 1.1}$	83.4$_{\pm 1.0}$	81.5$_{\pm 0.6}$	80.7$_{\pm 0.8}$	69.9$_{\pm 0.6}$	78.0$_{\pm 0.7}$
	DALR-RoBERTa	**73.1**$_{\pm 0.3}$	**83.2**$_{\pm 0.7}$	**76.5**$_{\pm 0.9}$	**83.9**$_{\pm 1.0}$	**82.2**$_{\pm 0.4}$	**81.2**$_{\pm 1.1}$	**72.0**$_{\pm 0.7}$	**78.9**$_{\pm 0.8}$

Table 1: Sentence representation performance on STS tasks (Spearman's correlation, "all" setting). Avg.: average performance across 7 tasks. ♡: results from (Gao et al., 2021), †: results from (Zhang et al., 2022a), ‡: results from (Nguyen et al., 2024). We train the models using different seeds and present the average and standard deviations of our findings. We highlight the highest numbers among models with the same pre-trained encoder.

total size N_t) and the multimodal dataset (with total size N_m) are different, we employed a mixed alternating sampling training strategy. Specifically, each epoch contains the total data from both datasets. By setting the ratio $N_t//N_m = a$, we load the data as follows: first, we load batches of pure-text data, followed by one batch of multimodal data. In each batch, the model's loss is updated. We evaluate on the development set of STS-B every 125 steps during training and retain the best checkpoint. All experiments are performed on a NVIDIA Tesla A100 (80GB) GPU. More training details can be found in Appendix B.

4.2 Main Results

Results on STS Tasks Table 1 reports the average STS results over five runs with different random seeds. It is clear that DALR significantly outperforms the previous methods on all PLMs. For example, in the *wiki+flickr* setting, compared with KDMCSE, DALR improves $\text{BERT}_{\text{base}}$ from 78.6% to 79.5% (+0.9%) and $\text{RoBERTa}_{\text{base}}$ from 79.1% to 79.9% (+0.8%). Compared to previous state-of-the-art methods, DALR still achieves consistent improvements, demonstrating that DALR provides stronger discriminative representations on the STS tasks. These results also dedicate the effectiveness of our approach in leveraging visual information to boost text representation learning.

Results on TR Tasks We train a logistic regression classifier under the premise of freezing the sentence embedding and evaluate its classification accuracy. As shown in Table 2, the experimental results show that our method achieves the best performance across all tasks on all PLMs, and the overall performance is better than other baselines. Specifically, compared to MCSE, our method achieves absolute improvements of 1.28% and 1.16% on the *wiki+flickr* dataset. On the *wiki+coco* dataset, our approach increases performance from 85.46% to 86.61% with BERT and from 85.85% to 86.73% with RoBERTa. This further verifies the effectiveness of our method in the transfer tasks.

4.3 Ablation Studies

To validate the effectiveness and necessity of the proposed strategies in DALR, we conduct ablation studies using the $\text{BERT}_{\text{base}}$ on the mixed "*wiki+flickr*" dataset. As shown in Table 3, when cross-modal alignment learning (CML) is removed, the performance drops significantly across all metrics. This highlights the importance of CML, indicating that incorporating knowledge from other modalities helps in learning more comprehensive representations. A similar degradation is observed when intra-modal alignment learning (IML) is removed, which demonstrates that IML effectively captures fine-grained semantic information and fa-

	Model	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	Avg.↑
wiki	SimCSE-BERT♡	82.92	87.23	95.71	88.73	86.81	87.01	78.07	86.64
	SimCSE-RoBERTa♡	83.37	87.76	95.05	87.16	89.02	90.80	75.13	86.90
wiki+flickr	MCSE-BERT♢	82.07	87.28	94.96	89.61	86.58	84.04	74,93	85.64
	KDMCSE-BERT♢	82.78	87.89	95.37	90.08	87.61	86.08	75.88	86.53
	DALR-BERT	**82.95**	**88.10**	**95.89**	**90.83**	**88.04**	**86.60**	**76.06**	**86.92**
	MCSE-RoBERTa♢	82.82	88.04	95.70	90.13	87.09	84.97	75.51	86.29
	KDMCSE-RoBERTa♢	83.21	88.16	95.73	90.46	88.05	86.30	76.18	86.87
	DALR-RoBERTa	**83.57**	**88.69**	**96.44**	**91.01**	**88.96**	**86.80**	**76.74**	**87.45**
wiki+coco	MCSE-BERT♢	81.75	86.89	94.73	89.44	86.81	83.97	74,66	85.46
	KDMCSE-BERT♢	82.30	87.71	95.04	89.86	87.38	85.68	75.51	86.20
	DALR-BERT	**82.66**	**87.90**	**95.85**	**90.43**	**87.59**	**86.09**	**75.74**	**86.61**
	MCSE-RoBERTa♢	82.24	87.53	95.22	89.76	87.08	84.15	74.96	85.85
	KDMCSE-RoBERTa♢	82.47	87.88	95.24	89.95	87.51	85.77	75.82	86.37
	DALR-RoBERTa	**82.71**	**88.02**	**96.10**	**90.21**	**87.85**	**86.38**	**75.84**	**86.73**

Table 2: Transfer task results of different sentence representation models (measured as accuracy). Avg.: average across 7 tasks. ♡: results from (Gao et al., 2021); ♢: reproduce the models (Zhang et al., 2022a; Nguyen et al., 2024) based on publicly available code. We highlight the highest numbers among models with the same PLM.

		STS (Avg.) ↑	TR (Avg.) ↑
wiki+flickr	**DALR**	**79.49** $_{\pm 0.7}$	**86.92** $_{\pm 1.0}$
	w/o $\mathcal{L}_{Info}$	78.24 $_{\pm 0.9}$	85.95 $_{\pm 0.4}$
	w/o $\mathcal{L}_{CML}$	78.16 $_{\pm 0.8}$	85.72 $_{\pm 1.1}$
	w/o $\mathcal{L}_{consistency}$	79.15 $_{\pm 0.7}$	86.70 $_{\pm 0.9}$
	w/o $\mathcal{L}_{CMA}$	78.61 $_{\pm 0.3}$	86.22 $_{\pm 0.5}$
	w/o $\mathcal{L}_{IML}$	78.82 $_{\pm 0.4}$	86.43 $_{\pm 0.7}$
	w/o $\mathcal{L}_{rank}$	79.06 $_{\pm 0.6}$	86.63 $_{\pm 1.0}$
	w/o $\mathcal{L}_{IMA}$	78.91 $_{\pm 0.8}$	86.50 $_{\pm 1.1}$
	w/o $\mathcal{L}_{IML}$&$\mathcal{L}_{CML}$	77.17 $_{\pm 0.7}$	85.54 $_{\pm 0.5}$

Table 3: Ablation study on our train loss. We quantify the individual contributions of the components: traditional multimodal contrastive loss ($\mathcal{L}_{Info}$), cross-modal alignment loss ($\mathcal{L}_{CML}$), and intra-modal alignment loss ($\mathcal{L}_{IML}$) (reported avg and std over 5 runs).

Model	image → text		text → image	
	R@1	R@5	R@1	R@5
MCSE†	16.7	43.5	22.5	50.4
KDMCSE†	17.9	45.0	24.1	52.8
w/ CML†	19.1	46.4	25.6	54.0
DALR†	**19.5**	**47.6**	**26.7**	**55.9**
MCSE‡	8.8	26.6	10.9	31.2
KDMCSE‡	9.4	27.9	12.2	32.7
w/ CML‡	9.7	28.6	13.3	33.9
DALR‡	**10.2**	**29.0**	**13.9**	**34.3**

Table 4: Multimodal retrieval results on Flickr30k test set based on $\text{BERT}_{\text{base}}$. † and ‡ denote the settings of *wiki+flickr* and *wiki+coco*, respectively.

cilitates the learning of more accurate and nuanced representations. Pairwise combinations of these components also yield noticeable improvements, highlighting the strength of our approach. Owing to the constraints of space, an in-depth exploration of experiments conducted on the "*wiki+coco*" dataset is meticulously detailed in Appendix E.1. Additionally, a comprehensive analysis of diverse teacher models is presented in Appendix E.2.

4.4 Analysis and Discussion

Components Analysis To verify the impact of cross-modal alignment (CML) in Eq.8, we integrate the CML into KDMCSE and evaluate its performance on retrieval tasks (details in Appendix E.3). As shown in Table 4, "KDMCSE + CML" outperforms "KDMCSE", demonstrating that while static threshold filtering reduces false negatives, it fails to fully address cross-modal biases. These biases arise from modality heterogeneity, and simple similarity thresholds are insufficient for aligning global semantic features across modalities.

For deeper analysis, we test intra-modal alignment (IML) on text-based tasks such as re-ranking, retrieval, and classification using the MTEB benchmark (Muennighoff et al., 2023). Table 5 shows that incorporating IML ("KDMCSE + IML") significantly improves performance, underscoring the importance of addressing ISD for better sentence representations.

Visualization Analysis To deeply assess the impact of each component effect, we conduct visualize experiments using $\text{BERT}_{\text{base}}$ with all components included and with specific components removed. We randomly sample 5,000 image-text pairs from the MSCOCO test set and generate their corresponding text embeddings. These embeddings

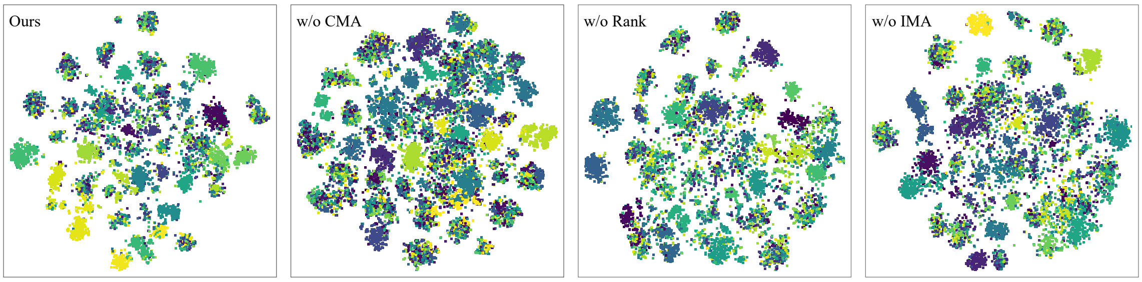

Figure 3: The t-SNE of sentence representations learned by DLAR and its three deviants (w/o specific component) using $\text{BERT}_{\text{base}}$. The points are embeddings of sentences sampled from the MSCOCO dataset(Xu et al., 2017). We use K-Means clustering to group similar sentence embeddings and form 50 clusters. (Best viewed in color)

are then projected into a lower-dimensional space using t-SNE (Reif et al., 2019), as shown in Figure 3. The visualization reveals that removing any component disrupts the clustering of similar sentence pairs (indicated by the same color), resulting in poor separation. Conversely, with all components jointly employed, similar samples are effectively clustered, while dissimilar samples remain well-separated. This highlights the ability of our method to improve semantic clustering and reduce representation bias.

Discussion with LLMs Sentence representation methods based on LLMs often rely on supervised signals, such as generating positive and negative samples (Wang et al., 2023; Li et al., 2024) or using instruction tuning (Cheng et al., 2023a), which may lead to unfair comparisons. For example, BGE (Xiao et al., 2024) asymmetrically adds scene descriptions to questions to improve generalization and trains with a large batch size of 19,200, significantly boosting performance. Our study focuses on enhancing sentence representations through images under an unsupervised paradigm similar to SimCSE. Unlike resource-intensive LLM-based approaches, our lightweight model is tailored for retrieval and ranking tasks, prioritizing efficiency and scalability. In many real-world applications, LLMs are impractical due to high computational costs and slower inference, making our method a more efficient and scalable alternative.

More Evaluation Metrics To validate the robustness and generalization ability of our method and scientifically include more diverse experimental evaluation metrics, we further evaluate its performance on additional downstream tasks. As shown in Table 5, our proposed method achieves superior performance compared to baseline models across multiple tasks, including reranking (Re-Rank), retrieval (Retrieval), and classification (CLF). Our

Model	**Re-Rank**	**CLF**	**Retrieval**	**STS**
SimCSE$^\heartsuit$	46.47	62.54	20.29	74.33
MCSE$^\diamondsuit$	46.92	63.20	21.43	77.02
KDMCSE$^\diamondsuit$	47.50	64.83	22.06	78.34
w/ IML	47.96	65.32	22.67	78.81
DALR (ours)	**48.35**	**67.46**	**23.84**	**79.38**
Δ	+0.85	+2.63	+1.78	+1.04

Table 5: Downstream tasks performance among our method and baselines on $\text{BERT}_{\text{base}}$ using *wiki+flickr*. $\heartsuit$: results from (Muennighoff et al., 2023), $\diamondsuit$: reproduce the models based on publicly available code.

Model	***Alignment*** $\downarrow$		***Uniformity*** $\downarrow$	
	flickr	*coco*	*flickr*	*coco*
MCSE-BERT	0.293	0.267	**-2.491**	-2.350
KDMCSE-BERT	0.245	0.261	-2.387	-2.383
DALR-BERT	**0.178**	**0.247**	-2.215	**-2.390**
MCSE-RoBERTa	0.209	0.195	-1.721	-1.418
KDMCSE-RoBERTa	0.174	0.149	-1.952	-1.748
DALR-RoBERTa	**0.153**	**0.136**	**-1.977**	**-1.785**

Table 6: The alignment uniformity results of the models when using BERT and RoBERTa. All models are trained in the *wiki-flickr* setting.

comprehensive evaluations not only substantiate the effectiveness of our approach but also guarantee a diverse and exhaustive performance assessment.

Alignment and Uniformity Prior work (Wang and Isola, 2020) has demonstrated that models with better *alignment* and *uniformity* can achieve better performance (detailed in Appendix D). We calculate the alignment and uniformity loss on the STS-B development set every 125 training steps. As shown in Table 6, compared to the previous baseline methods, DALR demonstrates superior performance in both *alignment* and *uniformity*, particularly in alignment. This indicates that our alignment strategies significantly enhance the alignment of sentence embeddings, thereby improving the overall quality of the embeddings. To further verify our results, we also conduct experiments on eliminating anisotropy (detailed in Appendix F).

5 Conclusion

In this paper, we propose a dual-level alignment framework (DALR) for multimodal sentence representation learning. DALR extends traditional multimodal contrastive learning by promoting both cross-modal and intra-modal alignment for more robust sentence representations. We introduce an auxiliary task to refine negative sampling and generate similarity matrices for effective cross-modal alignment. Intra-modal alignment is achieved through a combination of ranking distillation and KL divergence-based fine-grained calibration. Extensive experiments on STS and TR benchmarks, supported by detailed analyses, show that DALR consistently outperforms previous state-of-the-art methods.

Limitations

In this paper, the limitations of our work are as follows. Firstly, there are significant differences in the word token distributions and sizes between image-text datasets like MSCOCO and Flickr30k and traditional language corpora (e.g., Wikipedia). While Wikipedia contains billions of words, MSCOCO only contains about 1 million words. Empirically, performance improves with more training data. Secondly, building sentence representation models suited for few-shot learning is a key direction for future research, especially in scenarios where collected data is scarce.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62176187).

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43.

Patrick Bordes, Eloi Zablocki, Laure Soulier, Benjamin Piwowarski, and Patrick Gallinari. 2020. Incorporating visual semantics into sentence representations within a grounded space. *arXiv preprint arXiv:2002.02734*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14.

Qinyuan Cheng, Xiaogui Yang, Tianxiang Sun, Linyang Li, and Xipeng Qiu. 2023a. Improving contrastive learning of sentence embeddings from ai feedback. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11122–11138.

Xize Cheng, Tao Jin, Linjun Li, Wang Lin, Xinyu Duan, and Zhou Zhao. 2023b. OpenSR: Open-modality speech recognition via maintaining multi-modality alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6592–6607.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. 2022. DiffCSE: Difference-based contrastive learning for sentence embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4207–4218.

Sanghyuk Chun, Wonjae Kim, Song Park, Minsuk Chang, and Seong Joon Oh. 2022. Eccv caption: Correcting false negatives by collecting machine-and-human-verified image-caption associations for ms-coco. In *European Conference on Computer Vision*, pages 1–19. Springer.

Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. 2021. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8415–8424.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Jinghao Deng, Fanqi Wan, Tao Yang, Xiaojun Quan, and Rui Wang. 2023. Clustering-aware negative sampling for unsupervised sentence representation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8713–8729.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. 2024. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26584–26595.

Hongliang He, Junlei Zhang, Zhenzhong Lan, and Yue Zhang. 2023. Instance smoothed contrastive learning for unsupervised sentence embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12863–12871.

Kang He, Yuzhe Ding, Bobo Li, Haining Wang, Fei Li, Chong Teng, and Donghong Ji. 2025. Harnessing dimensional contrast and information compensation for sentence embedding enhancement. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1367–1377.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Jian Huang, Yanli Ji, Yang Yang, and Heng Tao Shen. 2023a. Cross-modality representation interactive learning for multimodal sentiment analysis. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 426–434.

Jingjia Huang, Yinan Li, Jiashi Feng, Xinglong Wu, Xiaoshuai Sun, and Rongrong Ji. 2023b. Clover: Towards a unified video-language alignment and fusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14856–14866.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. *Advances in neural information processing systems*, 28.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Ce Liu, and Lijuan Wang. 2023. Lavender: Unifying video-language understanding as masked language modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23119–23129.

Mingxin Li, Richong Zhang, Zhijie Nie, and Yongyi Mao. 2024. Narrowing the gap between supervised and unsupervised sentence representation learning with large language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13590–13599.

Zejun Li, Zhihao Fan, Huaixiao Tou, Jingjing Chen, Zhongyu Wei, and Xuanjing Huang. 2022. Mvptr: Multi-level semantic alignment for vision-language pre-training via multi-stage learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4395–4405.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755.

Yan Ling, Jianfei Yu, and Rui Xia. 2022. Vision-language pre-training for multimodal aspect-based sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2149–2159.

Jiduan Liu, Jiahao Liu, Qifan Wang, Jingang Wang, Wei Wu, Yunsen Xian, Dongyan Zhao, Kai Chen, and Rui Yan. 2023a. RankCSE: Unsupervised sentence representations learning via learning to rank. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13785–13802.

Ye Liu, Lingfeng Qiao, Changchong Lu, Di Yin, Chen Lin, Haoyuan Peng, and Bo Ren. 2023b. Osan: A one-stage alignment network to unify multimodal alignment and unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3551–3560.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A sick cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037.

Cong-Duy Nguyen, Thong Nguyen, Duc Vu, and Anh Luu. 2023. Improving multimodal sentiment analysis: Supervised angular margin-based contrastive learning for enhanced fusion representation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14714–14724.

Cong-Duy Nguyen, Thong Nguyen, Xiaobao Wu, and Anh Tuan Luu. 2024. KDMCSE: Knowledge distillation multimodal sentence embeddings with adaptive angular margin contrastive learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 733–749.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 271–278.

Bo Pang and Lillian Lee. 2005. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, page 115–124.

Zarana Parekh, Jason Baldridge, Daniel Cer, Austin Waters, and Yinfei Yang. 2021. Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for ms-coco. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2855–2870.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of bert. *Advances in Neural Information Processing Systems*, 32.

Yeon Seonwoo, Guoyin Wang, Changmin Seo, Sajal Choudhary, Jiwei Li, Xiang Li, Puyang Xu, Sunghyun Park, and Alice Oh. 2023. Ranking-enhanced unsupervised sentence representation learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15783–15798.

Zhan Shi, Guoyin Wang, Ke Bai, Jiwei Li, Xiang Li, Qingjun Cui, Belinda Zeng, Trishul Chilimbi, and Xiaodan Zhu. 2023. Osscse: Overcoming surface structure bias in contrastive learning for unsupervised sentence embedding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7242–7254.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

Kim Sung-Bin, Arda Senocak, Hyunwoo Ha, Andrew Owens, and Tae-Hyun Oh. 2023. Sound to visual scene generation by audio-to-visual latent alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6430–6440.

Zineng Tang, Jaemin Cho, Hao Tan, and Mohit Bansal. 2021. Vidlankd: Improving language understanding via video-distilled knowledge transfer. *Advances in Neural Information Processing Systems*, 34:24468–24481.

Zhiliang Tian, Zheng Xie, Fuqiang Lin, and Yiping Song. 2023. A multi-view meta-learning approach for multi-modal response generation. In *Proceedings of the ACM Web Conference 2023*, pages 1938–1947.

Ellen M Voorhees and Dawn M Tice. 2000. Building a question answering test collection. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 200–207.

Haining Wang, Kang He, Bobo Li, Lei Chen, Fei Li, Xu Han, Chong Teng, and Donghong Ji. 2024. Refining and synthesis: A simple yet effective data augmentation framework for cross-domain aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10318–10329.

Qian Wang, Weiqi Zhang, Tianyi Lei, Yu Cao, Dezhong Peng, and Xu Wang. 2023. Clsep: Contrastive learning of sentence embedding with prompt. *Knowledge-Based Systems*, 266:110381.

Qifan Wang, Yi Fang, Anirudh Ravula, Fuli Feng, Xiaojun Quan, and Dongfang Liu. 2022a. Webformer: The web-page transformer for structure information extraction. In *Proceedings of the ACM Web Conference 2022*, pages 3124–3133.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of International conference on machine learning*, pages 9929–9939.

Weizhi Wang, Li Dong, Hao Cheng, Haoyu Song, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2022b. Visually-augmented language modeling. *arXiv preprint arXiv:2205.10178*.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39:165–210.

Qiyu Wu, Chongyang Tao, Tao Shen, Can Xu, Xiubo Geng, and Daxin Jiang. 2022a. PCL: Peer-contrastive learning with diverse augmentations for unsupervised sentence embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 12052–12066.

Xing Wu, Chaochen Gao, Liangjun Zang, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2022b. ESimCSE: Enhanced sample building method for contrastive learning of unsupervised sentence embedding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3898–3907.

Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. 2008. Listwise approach to learning to rank: theory and algorithm. In *Proceedings of the 25th international conference on Machine learning*, pages 1192–1199.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.

Jiaming Xu, Bo Xu, Peng Wang, Suncong Zheng, Guanhua Tian, and Jun Zhao. 2017. Self-taught convolutional neural networks for short text clustering. *Neural Networks*, 88:22–31.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5065–5075.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. 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.

Tianshu Yu, Haoyu Gao, Ting-En Lin, Min Yang, Yuchuan Wu, Wentao Ma, Chao Wang, Fei Huang, and Yongbin Li. 2023. Speech-text pre-training for spoken dialog understanding with explicit cross-modal alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7900–7913.

Chunhui Zhang, Xin Sun, Yiqian Yang, Li Liu, Qiong Liu, Xi Zhou, and Yanfeng Wang. 2023. All in one: Exploring unified vision-language tracking with multi-modal alignment. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 5552–5561.

Miaoran Zhang, Marius Mosbach, David Adelani, Michael Hedderich, and Dietrich Klakow. 2022a. MCSE: Multimodal contrastive learning of sentence embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5959–5969.

Yanzhao Zhang, Richong Zhang, Samuel Mensah, Xudong Liu, and Yongyi Mao. 2022b. Unsupervised sentence representation via contrastive learning with mixing negatives. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11730–11738.

Li Zheng, Boyu Chen, Hao Fei, Fei Li, Shengqiong Wu, Lizi Liao, Donghong Ji, and Chong Teng. 2024. Self-adaptive fine-grained multi-modal data augmentation for semi-supervised muti-modal coreference resolution. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8576–8585.

Li Zheng, Hao Fei, Ting Dai, Zuquan Peng, Fei Li, Huisheng Ma, Chong Teng, and Donghong Ji. 2025. Multi-granular multimodal clue fusion for meme understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26057–26065.

Kun Zhou, Beichen Zhang, Xin Zhao, and Ji-Rong Wen. 2022. Debiased contrastive learning of unsupervised sentence representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6120–6130.

Minghao Zhu, Xiao Lin, Ronghao Dang, Chengju Liu, and Qijun Chen. 2023. Fine-grained spatiotemporal motion alignment for contrastive video representation learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 4725–4736.

Wenjie Zhuo, Yifan Sun, Xiaohan Wang, Linchao Zhu, and Yi Yang. 2023. WhitenedCSE: Whitening-based contrastive learning of sentence embeddings. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12135–12148.

A Baselines Model

We introduce a classic sentence embedding model and two typical multimodal sentence embedding models, which we implement using official code:

- SimCSE (Gao et al., 2021): conducts thorough experiments in both unsupervised and supervised settings using different dropout to obtain positive pairs.
- MCSE (Zhang et al., 2022a): introduces visual information in sentence embedding to enhance SimCSE, and captures the consistency of sentences and their related images in the same space.
- KDMCSE (Nguyen et al., 2024): inherits the knowledge of the teacher model to learn the distinction between positive and negative samples, while also proposing an adaptive angular margin supervised contrastive learning approach to enhance discriminability by reinforcing margins in the angular space.

B Implementation

Teacher Image Model We employ CLIP as our teacher model, which leverages contrastive learning to derive general visual and language representations from large-scale image-text pairs. The pre-trained weights are loaded from CLIP-ViT-B/32, with the patch size set to 32. After loading the model to obtain the image features, we feed them into a MLP for projection into a shared 256-dimensional space.

Teacher Text Model We propose using a multi-teacher model weighting strategy to obtain the final teacher representations. In this work, we follow the same setup as RankCSE (Liu et al., 2023a), utilizing SimCSE (Gao et al., 2021) and DiffCSE (Chuang et al., 2022) as teacher models, and the final teacher representation is obtained through weighted aggregation. Additionally, the feature representations are projected into a shared 256-dimensional space. Moreover, other text teacher models such as RankCSE and CLIP (Radford et al., 2021) can also be substituted. A detailed comparison is provided in Appendix E.2.

Student Language Model The implementation of the language encoder is based on the Transformers library. We start with the checkpoints of bert-base-uncased and roberta-base, fine-tuning

	KDMCSE		DALR	
	wiki+flickr	*wiki+coco*	*wiki+flickr*	*wiki+coco*
Batch size	128	128	128	128
Epoch	4	4	4	4
Total time	290 min	440 min	245 min	370 min

Table 7: Training Efficiency of KDMCSE and DALR based on $\text{BERT}_{\text{base}}$.

the pre-trained models using our proposed training objective in Eq.14. For evaluation, we use the 768-dimensional [CLS] token output prior to MLP pooling layer as the sentence embedding. For the MLP projection head, in the plain text setting (using the Wiki1M dataset), the sentence embeddings are projected into a 768-dimensional space. In the multimodal setting, the feature representations are projected into a shared 256-dimensional space.

More Implementation Details We preform experiments with backbones of $\text{BERT}_{\text{base}}$ and $\text{RoBERTa}_{\text{base}}$. We choose [CLS] embeddings as the final representation. In the plain text setting (using Wiki1M), sentence representations are projected into a 768-dimensional space. In the multimodal setting, the student and teacher models' feature representations are projected into a shared 256-dimensional space. We use two mixed text and multimodal training scenarios: *wiki+flickr* and *wiki+coco*. We evaluate on the development set of STS-B every 125 steps during training and retain the best checkpoint. We implement all experiments with the deep learning framework PyTorch on a NVIDIA Tesla A100 GPU (80GB memory). The temperature parameter τ is set to 0.05, and the weight parameters λ and μ are set to 0.1 and 0.2, respectively. For $\text{BERT}_{\text{base}}$ encoder, we use a learning rate of 2e-5 and a batch size of 128 for training; for $\text{RoBERTa}_{\text{base}}$, the learning rate is 1e-5 and the batch size is also set to 128. The runtime for each of our experiments is approximately 4 hours, which is shorter than KDMCSE. More details are provided in Appendix C.

C Training Efficiency

We compare the training efficiency of KDMCSE and DALR using $\text{BERT}_{\text{base}}$, both tested on a single NVIDIA Tesla A100 GPU (with 80GB of memory). In the experiments, we set the batch size of KDMCSE and DALR to 128, and the training epochs to 4. As shown in Table 7, under the *wiki+flickr* and *wiki+coco* experimental settings, DALR completes training in 4 hours and 6.2 hours, respectively.

		STS (Avg.) ↑	TR (Avg.) ↑
wiki+coco	**DALR**	**78.64**$_{\pm 0.9}$	**86.61**$_{\pm 0.7}$
	w/o $\mathcal{L}_{Info}$	77.20$_{\pm 1.0}$	85.39$_{\pm 0.8}$
	w/o $\mathcal{L}_{CML}$	77.48$_{\pm 0.6}$	85.53$_{\pm 0.8}$
	w/o $\mathcal{L}_{consistency}$	78.19$_{\pm 0.4}$	86.42$_{\pm 0.7}$
	w/o $\mathcal{L}_{CMA}$	77.85$_{\pm 0.5}$	85.70$_{\pm 0.9}$
	w/o $\mathcal{L}_{IML}$	77.89$_{\pm 0.8}$	85.74$_{\pm 0.5}$
	w/o $\mathcal{L}_{rank}$	78.31$_{\pm 1.1}$	86.45$_{\pm 0.7}$
	w/o $\mathcal{L}_{IMA}$	78.07$_{\pm 0.9}$	86.33$_{\pm 1.0}$
	w/o $\mathcal{L}_{IML}$&$\mathcal{L}_{CML}$	76.75$_{\pm 0.6}$	85.07$_{\pm 1.2}$

Table 8: Ablation study on our train loss based on *wiki+coco*. We quantify the individual contributions of the components: traditional multimodal contrastive loss ($\mathcal{L}_{Info}$), cross-modal alignment loss ($\mathcal{L}_{CML}$), and intra-modal alignment loss ($\mathcal{L}_{IML}$) (reported avg and std over 5 runs).

D Alignment and Uniformity

Contrastive representation learning has two key properties: (1) *alignment* of positive pairs; (2) *uniformity* on the hypersphere. Wang and Isola (2020) argues that directly optimizing these two metrics can lead to representations with performance comparable to or better than contrastive learning in downstream tasks. *Alignment* measures the expected distance between normalized representations of positive pairs p_{pos}:

$$\ell_{\text{align}} \triangleq \mathop{\mathbb{E}}_{(x,x^+)\sim p_{\text{pos}}} \|f(x) - f(x^+)\|^2, \tag{15}$$

while *uniformity* measures the uniform distribution of normalized representations:

$$\ell_{\text{uniform}} \triangleq \log \mathop{\mathbb{E}}_{x,y \overset{i.i.d.}{\sim} p_{\text{data}}} e^{-2\|f(x)-f(y)\|^2}, \tag{16}$$

where p_{data} represents the distribution of sentence pairs. Smaller values for both metrics are better, which aligns closely with the objectives of contrastive learning: positive instances should be as close as possible, indicating smaller alignment, while random instances should be scattered on the hypersphere, indicating smaller uniformity.

E Analysis

E.1 More Ablation studies

Due to space constraints, we present the ablation study results on *wiki+coco* here. The results in Table 8 demonstrate that all three components are essential, as the absence of any of them leads to a performance drop. Notably, the cross-modal alignment module has the most significant impact on

Teacher Model		STS(Avg.)	TR(Avg.)
Image	Text		
CLIP	SimCSE	77.84	85.17
CLIP	DiffCSE	78.65	86.23
CLIP	CLIP	79.42	86.89
CLIP	SimCSE+DiffCSE	79.49	86.92
CLIP	SimCSE+RankCSE	79.61	87.05
ResNet	SimCSE	77.59	85.03
ResNet	DiffCSE	78.28	85.97
ResNet	CLIP	79.04	86.45
ResNet	SimCSE+DiffCSE	79.32	86.76
ResNet	SimCSE+RankCSE	79.36	86.80

Table 9: Comparisons of different image and text teachers based on *wiki+flickr* setting using $\text{BERT}_{\text{base}}$.

performance, as it effectively leverages image information to provide supervisory signals for text representation learning.

E.2 Teacher Model Selection

We conduct extensive experiments to explore the impact of different teacher models (image and text) on DALR's performance. As illustrated in Figure 9, the results show that combining both cross-modal and intra-modal information can generate more discriminative sentence representations. By comparing various teacher models, we found that stronger teacher models lead to improved performance, which aligns with our expectations. ResNet is trained solely on image data and lacks multi-modal capabilities. As a result, when used as an image teacher model, the sentence representations it helps learn tend to be slightly less effective. A more powerful image teacher model can capture finer details of visual information, while a more advanced text teacher model provides more accurate ranking labels, facilitating more precise ranking knowledge transfer. We also observe an interesting phenomenon: using SimCSE and RankCSE as teacher models yielded even better results than those in our main experiments in Section 4.2. This suggests that further investigation into the selection of teacher models could provide valuable insights for future research.

E.3 Cross-modal Retrieval

To comprehensively evaluate the performance of cross-modal retrieval, we use the R@k metric as the standard for assessing cross-modal retrieval datasets. DALR is designed to learn high-quality sentence embeddings, with a primary focus on semantic similarity tasks. However, its integration of cross-modal contrastive learning and alignment modules also enhances performance in cross-modal retrieval. This outstanding performance further validates the effectiveness and robustness of our model.

F Anisotropy Study

Recent research (Ethayarajh, 2019) has highlighted the anisotropy issue in language representations, wherein learned embeddings are confined to a narrow cone in vector space, severely restricting their expressive capacity. Specifically, the anisotropy in sentence representations results in vectors being densely clustered in specific directions, diminishing their ability to effectively distinguish between different sentences.

To evaluate the impact of our method on mitigating anisotropy, we display the cosine similarity between sentence pair representations calculated on the STS-B test set, and compare them with the gold-standard annotations on STS-B. The Y-axis represents the cosine similarity of the sentence pairs, while the X-axis corresponds to the annotation scores (ranging from 0 to 5), with higher annotation scores indicating greater similarity. In other words, for sentence pairs annotated with a score of 5, the computed cosine similarity should be high. Each light-colored dot represents a sentence pair, and due to the large number of samples, overlapping dots may appear darker.

As shown in Figure 4, the results demonstrate that, for low-scoring sentence pairs, the predicted similarity by our model is significantly lower, outperforming the SimCSE, MCSE, and KDMCSE methods. This outcome also indicates that the anisotropy issue has been alleviated to some extent.

G Qualitative Analysis

We conduct small-scale retrieval experiments using KDMCSE and DLAR based on $\text{BERT}_{\text{base}}$. We use 30k captions from the Flickr30k (Young et al., 2014) dataset as the retrieval data and randomly select any sentence from them as a query to retrieve the Top-3 similar sentences (based on cosine similarity). As shown in Table 10, the retrieval results demonstrate that sentences retrieved by DLAR are semantically closer to the query sentences and of higher quality compared to those retrieved by DKMCSE, further demonstrating the effectiveness of DALR.

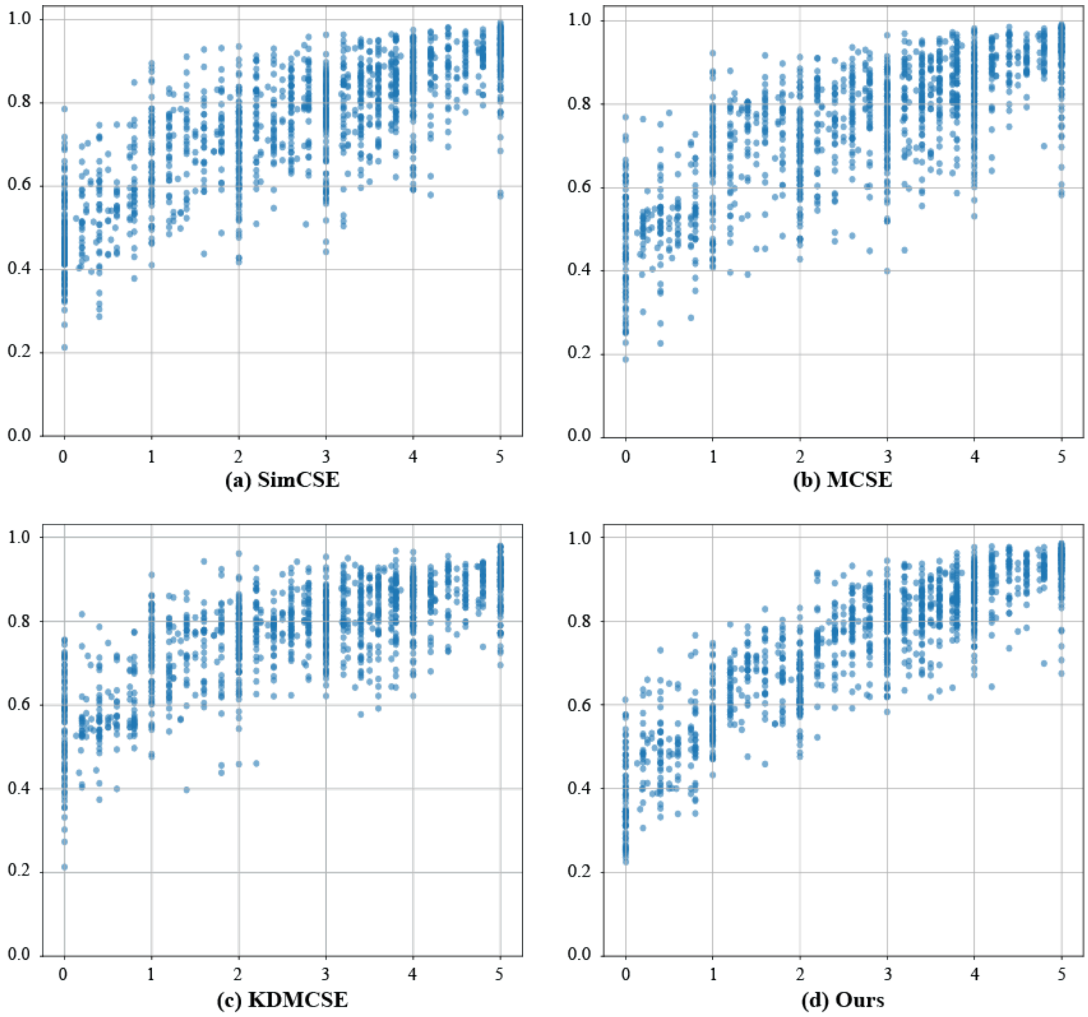

Figure 4: Scatter plot of the ground truth similarity scores (x-axis) and the cosine similarities (y-axis) between sentence pairs in the STS-B (test set). Each entry in the STS-B includes a text pair and a similarity score from 0 to 5 (gold standard).

Rank	**KDMCSE**	**DALR (Ours)**
Query: A group of men climb ladders outdoors.		
#1	Two people standing on a roof while another climbs a ladder.	Two people standing on a roof while another climbs a ladder.
#2	A firefighter climbs a ladder towards the fire above him.	Two men sitting on the roof of a house while another one stands on a ladder.
#3	A person is climbing a wooden ladder up a rocky ledge.	Three people in t-shirt, yellow helmets and harnesses begin to climb ladder.
Query: A man in a white cap and shirt plays the violin with other street performers.		
#1	A man in a white shirt is playing the flute to someone in a red skirt.	A man in a white shirt is playing the flute to someone in a red skirt.
#2	A man in a white shirt plays an electric violin.	A man in a white shirt plays an electric violin.
#3	A man in a red shirt plays the guitar.	A man with glasses wearing a tie plays the violin.
Query: A man in a black outfit poses in front of the eiffel tower.		
#1	A man carrying trinkets with the Eiffel tower in the background.	A man carrying trinkets with the Eiffel tower in the background.
#2	A man wearing black jacket poses with a smile.	A man in formal wear is posing in front of a building.
#3	A man wearing a black long-sleeved shirt is taking a photo of a building.	A man wearing a black long-sleeved shirt is taking a photo of a building.
Query: Two women wearing ceremonial costumes are walking outside a white building.		
#1	Two women wearing blue jeans are walking outside.	Two women wearing dresses are walking by a building.
#2	Two women wearing dresses are walking by a building.	Two people are wearing flower costumes and walking down a street.
#3	Men in traditional dress stand outside .	Two women wearing skirts and heels walking down a sidewalk.

Table 10: Retrieval examples of retrieved Top-3 sentences from queries by KDMCSE and DLAR from Flickr30k dataset (30k sentences).