Improving Multimodal Contrastive Learning of Sentence Embeddings with Object-Phrase Alignment

Kaiyan Zhao, Zhongtao Miao, Yoshimasa Tsuruoka
The University of Tokyo, Tokyo, Japan
{kaiyan1006, miao, tsuruoka}@logos.t.u-tokyo.ac.jp

Abstract

Multimodal sentence embedding models typically leverage image-caption pairs in addition to textual data during training. However, such pairs often contain noise, including redundant or irrelevant information on either the image or caption side. To mitigate this issue, we propose MCSEO, a method that enhances multimodal sentence embeddings by incorporating fine-grained object-phrase alignment alongside traditional image-caption alignment. Specifically, MCSEO utilizes existing detection and segmentation models to extract accurate object-phrase pairs, which are then used to optimize a proposed contrastive learning objective tailored to object-phrase correspondence. Experimental results on semantic textual similarity (STS) tasks across different backbone models demonstrate that MCSEO outperforms image-caption-only baselines on average, and that it is complementary to existing methods: combining MCSEO with KDMCSE yields further gains. A random-pairing control and a manual evaluation of the extracted pairs additionally indicate that the improvements stem from correct object-phrase alignment.

1 Introduction

Sentence embedding aims to capture the semantic meaning of sentences by projecting them into fixed vectors in a shared high-dimensional space (Reimers and Gurevych, 2019; Gao et al., 2021), typically trained via contrastive learning (CL) (van den Oord et al., 2019). Recently, several works have incorporated visual context alongside textual information to learn *multimodal sentence embeddings*, most of which rely on image-caption pairs to align the two modalities (Zhang et al., 2022; Jian et al., 2022; Li et al., 2023; Nguyen et al., 2024).

Nguyen et al. (2024) observe that current image-caption datasets often contain noise that introduces misleading training signals, particularly when visually similar images are treated as negatives within the same batch. We further argue that even pairs labeled as *positive* can be substantially noisy. Figure 1 shows two examples from Flickr30k (Plummer et al., 2015) illustrating noise on both sides: in the left example, the caption's "driving through" misrepresents a car that is actually stopped with its door open; in the right example, the image includes elements (cameraman, trees) absent from the caption, yielding incomplete cross-modal alignment.

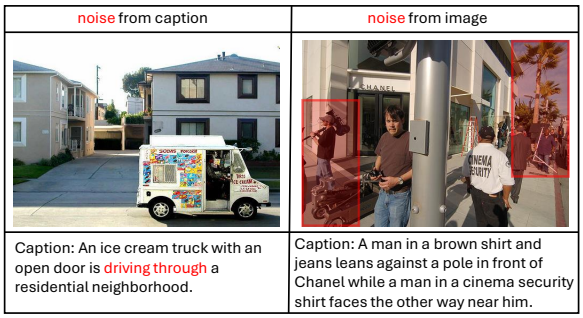


Figure 1: Image-caption pair examples from the Flickr30k dataset. We use red color to highlight the potential sources of noise. **Left**: The caption includes ambiguous or inaccurate details not supported by the image content. **Right**: The image contains complex visual elements that are not described in the caption.

Since such pairs serve as positives during training, directly minimizing the distance between noisy pairs can degrade performance. As perfectly aligned image–caption data is infeasible to guarantee at scale, we instead address the noise through *finer-grained* alignment: not only between images and captions, but also between objects and phrases. We propose MCSEO (Multimodal Contrastive learning of Sentence Embeddings with Object-phrase-level alignment), which leverages off-the-shelf segmentation and object detection models to extract objects from images and selectively decompose captions into correspond-

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 397–404
November 6-10, 2026 ©2026 Association for Computational Linguistics

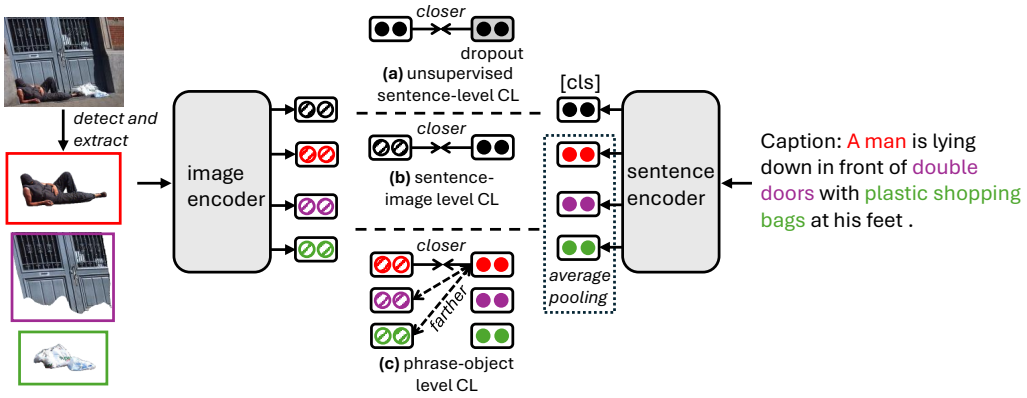


Figure 2: Overview of MCSEO. Given an image-caption pair, we decompose it into object-phrase pairs and then apply three training objectives designed to help the model capture more fine-grained multimodal information.

ing phrases, yielding cleaner object-phrase pairs. An object-phrase-level contrastive objective then guides the model to capture this alignment.

Experimental results across multiple backbone models and datasets demonstrate that MCSEO improves the average STS score over MCSE with statistical significance over five random seeds, and that the gains carry over to 39 MTEB (Muennighoff et al., 2023) tasks spanning classification, clustering, pair classification, retrieval and STS. MCSEO is also complementary to KDMCSE (Nguyen et al., 2024): combining the two yields the best results. A control experiment in which object–phrase correspondences are randomly shuffled falls below the MCSE baseline, and a manual evaluation shows 96.7% of extracted pairs are exactly correct, indicating that the gains come from accurate fine-grained alignment rather than from an additional loss term.

2 Related Works

Sentence embedding models have achieved strong performance with contrastive objective that pulls semantically similar sentences together and pushes dissimilar ones apart (Reimers and Gurevych, 2019; Gao et al., 2021; Zhao et al., 2024; Miao et al., 2024). A growing line of work extends this paradigm to *multimodal sentence embeddings* by incorporating non-linguistic signals, most commonly images (Zhang et al., 2022; Jian et al., 2022; Li et al., 2023; Nguyen et al., 2024). MCSE (Zhang et al., 2022) is an early attempt that uses an image encoder to align caption and image embeddings via CL, on top of standard textual training data. KDMCSE (Nguyen et al., 2024) examines noise in image-caption datasets, noting that visually distinct images are sometimes paired with near-identical captions, which yields misleading negatives.

Grounding textual phrases to image regions has been studied extensively, from datasets with manual region–phrase annotations such as Flickr30k (Plummer et al., 2015) to open-vocabulary grounding models such as MDETR (Kamath et al., 2021) and GLIP (Li et al., 2022b). Fine-grained region- or token-level alignment has also been used to improve vision–language pre-training (Li et al., 2022a). These works use fine-grained alignment to improve grounding, detection or cross-modal retrieval. We instead treat off-the-shelf grounding as a source of supervision and ask whether fine-grained alignment improves text-only sentence representations.

While KDMCSE targets noisy *negatives*, we address the complementary problem of improving *positive* pairs: rather than filtering or re-weighting whole pairs, we decompose each pair into cleaner object–phrase units, and we show empirically that the two approaches are complementary. Our contribution is thus not a new grounding method, but the application of fine-grained object–phrase alignment as a training signal for sentence embeddings.

3 Methods

An overview of our proposed MCSEO framework is illustrated in Figure 2. MCSEO employs three CL objectives during training: (a) unsupervised sentence-level CL, (b) sentence-image level CL, and (c) object-phrase level CL.

3.1 Unsupervised Sentence-level CL

Sentence embedding aims to train a model f_θ that maps a text $x_i \in \mathcal{D}_{text}$ to a fixed-size vector $h_i = f_\theta(x_i)$. In unsupervised CL, the posi-

tive example x_i^+ is typically obtained via dropout, while other sequences in the batch serve as negatives (Gao et al., 2021):

$$l_i^{text} = -\log \frac{e^{sim(h_i, h_i^+)/\tau}}{\sum_{j=1}^{N} e^{sim(h_i, h_j)/\tau}}, \quad (1)$$

where $sim(\cdot)$ is a similarity metric, τ is the temperature, and N is the batch size.

3.2 Sentence-image level CL

Given an image-caption pair $\{d_i^{img}, d_i^{cap}\}$ from a dataset $\mathcal{D}$, where i stands for the i_{th} example in the dataset, we can use an image encoder g_θ and a sentence encoder f_θ to obtain their respective representations:

$$h_i^{img} = g_\theta(d_i^{img}), \;\; h_i^{cap} = f_\theta(d_i^{cap}). \quad (2)$$

Since the caption describes the content of the corresponding image, it is natural to treat the paired image as the positive sample for the caption, while all other images in the batch serve as negative samples. The training objective for sentence-image level CL can be formulated as (Zhang et al., 2022):

$$l_i^{img-cap} = -\log \frac{e^{sim(h_i^{cap}, h_i^{img})/\tau'}}{\sum_{j=1}^{N} e^{sim(h_i^{cap}, h_j^{img})/\tau'}}, \quad (3)$$

where τ' is the temperature parameter.

As discussed earlier, directly aligning d_i^{img} and d_i^{cap} can be noisy due to imperfect correspondences. We therefore further refine the alignment by matching objects in the image with their corresponding phrases in the caption.

3.3 Extraction of Object-Phrase Pairs

We first apply a detection model f_{detect} capable of both segmentation and grounding (e.g., SAM2 (Ravi et al., 2025) with Florence2 (Xiao et al., 2024)) to decompose each image-caption pair into objects and phrases:

$$O_i, P_i = f_{detect}(\{d_i^{img}, d_i^{cap}\}), \quad (4)$$

where $O_i = \{o_i^1, o_i^2, ..., o_i^k\}$ and $P_i = \{p_i^1, p_i^2, ..., p_i^k\}$ represent the sets of detected objects and their matching phrases from the i_{th} image-caption pair. Note that each detected object is guaranteed to have a corresponding phrase, so both sets contain the same number of elements. Specifically, since using f_{detect} would output bounding boxes and labels, we instead use the detected masks for precisely extracting the objects in the given image.

We then encode the extracted objects and phrases:

$$h_o^{i,k} = g_\theta(o_i^k), \;\; h_p^{i,k} = f_\theta(p_i^k). \quad (5)$$

In practice, phrase embeddings are obtained by average-pooling the token embeddings of the full sentence at the corresponding phrase positions, avoiding redundant re-encoding while staying consistent with the sentence encoder.

3.4 Object-phrase level CL

For the i_{th} image-caption pair, we can conduct CL for the k_{th} phrase based on the detected objects and decomposed phrases from this pair:

$$l_i^{obj-phra} = -\sum_{p \in P_i, o \in O_i} \log \frac{e^{sim(h_p^{i,k}, h_o^{i,k})/\tau}}{\sum_{j=1}^{K} e^{sim(h_p^{i,k}, h_o^{i,j})/\tau}}, \quad (6)$$

where K is the number of elements in set O_i and P_i. This formulation encourages each phrase to be most similar to its corresponding object embedding compared to other objects. Note that this object-phrase-level loss is calculated within each image-caption pair instead of the whole batch.

The final loss of MCSEO is the combination of three CL losses:

$$l_i = l_i^{text} + \alpha l_i^{img-cap} + \beta l_i^{obj-phra}. \quad (7)$$

α and β are two hyperparameters that balance the contributions of each objective.

4 Experiments

4.1 Setup

We conduct experiments based on two kinds of text backbone models: BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). The `[CLS]` token is used as sentence embedding. For the image encoder, we use ResNet50 (He et al., 2016) to encode both images and segmented objects into a dimension of 2048. The image encoder is kept frozen during training and we only update the text encoder.

Following Gao et al. (2021) and Zhang et al. (2022), we use the Wikipedia subset as text-only dataset, Flickr30k (Plummer et al., 2015) and MSCOCO (Lin et al., 2014) as the image-caption datasets. To extract object-phrase alignments, we

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	avg.
text-only models (wiki)								
BERT	39.7	59.4	49.7	66.0	66.2	53.9	62.1	56.7
RoBERTa	40.9	58.7	49.1	65.6	61.5	58.6	61.6	56.6
SimCSE-BERT	68.4	82.4	74.4	80.9	78.6	76.9	72.2	76.3
SimCSE-RoBERTa	70.2	81.8	73.2	81.4	80.7	80.2	68.6	76.6
text-image models (wiki + Flickr)								
MCSE-BERT	71.4	81.8	74.8	83.6	77.5	79.5	72.6	77.3
MCSE-BERT (reproduced)	$71.6_{\pm 0.5}$	$81.3_{\pm 1.4}$	$75.0_{\pm 0.6}$	$83.8_{\pm 0.9}$	$77.4_{\pm 0.8}$	$79.9_{\pm 1.1}$	$72.8_{\pm 1.3}$	$77.4_{\pm 0.6}$
MCSEO-BERT (ours)	$71.1_{\pm 0.6}$	$82.5_{\pm 0.9}$	$75.7_{\pm 0.5}$	$84.6_{\pm 1.1}$	$80.1_{\pm 0.9}$	$81.4_{\pm 0.8}$	$72.9_{\pm 1.2}$	$\mathbf{78.3}^{\star}_{\pm 0.5}$
MCSE-RoBERTa	71.7	82.7	75.9	84.0	81.3	82.3	70.3	78.3
MCSE-RoBERTa (reproduced)	$72.9_{\pm 1.0}$	$82.8_{\pm 0.8}$	$75.5_{\pm 0.2}$	$83.2_{\pm 0.5}$	$79.9_{\pm 1.2}$	$82.8_{\pm 0.7}$	$71.1_{\pm 0.9}$	$78.3_{\pm 0.4}$
MCSEO-RoBERTa (ours)	$73.9_{\pm 0.4}$	$83.0_{\pm 0.7}$	$76.5_{\pm 0.5}$	$83.7_{\pm 0.3}$	$81.0_{\pm 1.0}$	$83.0_{\pm 0.9}$	$72.2_{\pm 0.4}$	$\mathbf{79.1}^{\star}_{\pm 0.6}$
text-image models (wiki + COCO)								
MCSE-BERT	71.2	79.7	73.8	83.0	77.8	78.5	72.1	76.6
MCSE-BERT (reproduced)	$71.8_{\pm 0.7}$	$79.8_{\pm 0.5}$	$74.1_{\pm 0.4}$	$82.5_{\pm 0.8}$	$78.0_{\pm 0.5}$	$79.3_{\pm 0.9}$	$71.9_{\pm 0.8}$	$76.8_{\pm 0.4}$
MCSEO-BERT (ours)	$72.1_{\pm 0.5}$	$81.1_{\pm 1.0}$	$74.9_{\pm 0.3}$	$83.3_{\pm 0.9}$	$80.2_{\pm 0.7}$	$80.1_{\pm 0.6}$	$73.4_{\pm 0.5}$	$\mathbf{77.9}^{\star}_{\pm 0.6}$
MCSE-RoBERTa	70.2	82.0	75.5	83.0	81.5	80.8	69.9	77.6
MCSE-RoBERTa (reproduced)	$70.8_{\pm 0.9}$	$82.0_{\pm 0.4}$	$75.4_{\pm 0.6}$	$82.9_{\pm 0.5}$	$81.1_{\pm 1.0}$	$81.3_{\pm 0.9}$	$70.0_{\pm 0.7}$	$77.6_{\pm 0.6}$
MCSEO-RoBERTa (ours)	$72.9_{\pm 0.5}$	$82.5_{\pm 0.7}$	$76.1_{\pm 0.7}$	$83.2_{\pm 0.5}$	$82.4_{\pm 0.6}$	$82.9_{\pm 0.9}$	$71.9_{\pm 0.6}$	$\mathbf{78.8}^{\star}_{\pm 0.7}$

Table 1: Evaluation results on seven STS tasks. We report Spearman's correlation scores for all tasks. Rows marked with "reproduced" and "ours" indicate mean results over five seeds; other results are quoted from Zhang et al. (2022). * denotes p<0.05 vs. the corresponding reproduced MCSE baseline (t-test on average). Best averages in bold.

use SAM2 with Florence2[1] as the detection and grounding model. All models are trained with a batch size of 64, and hyperparameters are tuned using STS-B's development set, with validation performed every 125 steps.

For evaluation, we select seven standard Semantic Textual Similarity (STS) tasks from the SentEval toolkit (Conneau and Kiela, 2018) to assess sentence embedding quality. Results are averaged over five random seeds. We additionally evaluate on 39 English MTEB tasks (Section 4.5).

Detailed setup can be found in Appendix A.1.

4.2 Experimental Results

The experimental results are presented in Table 1. We compare three categories of models: (1) *text-only* baselines trained solely on Wikipedia, (2) MCSE variants trained on Wikipedia and image-caption datasets (Flickr or COCO), and (3) our proposed method, MCSEO, which introduces additional object-phrase-level alignment during training. To ensure a fair comparison with MCSE, we also reproduce MCSE using the longest caption available for each image, consistent with the setup in MCSEO.

From Table 1, we observe several key findings. Firstly, object-phrase contrastive learning improves performance: MCSEO raises the average STS score over reproduced MCSE in all four settings, with all differences significant (p<0.05). Gains are consistent on average but not on every individual task. The improvements hold across both backbones and both image-caption datasets.

Notably, both MCSE and MCSEO are trained on the exact same datasets: Wikipedia and Flickr30k or COCO. The only difference is that MCSEO models additionally capture localized object-phrase alignment. The performance gains observed for MCSEO suggest that the quality of supervision (i.e., finer alignment) plays a crucial role, even when the quantity of data remains unchanged.

Objective	Pairing	STS avg.
l^{text}	-	76.3
$l^{text} + l^{img-cap}$ (MCSE)	-	77.4
$l^{text} + l^{obj-phra}$	correct	77.5
$l^{text} + l^{img-cap} + l^{obj-phra}$	random	76.3
$l^{text} + l^{img-cap} + l^{obj-phra}$ (MCSEO)	correct	78.3

Table 2: Ablation study on the loss function components. "Pairing" indicates whether object-phrase correspondences are correct or randomly shuffled.

4.3 Ablation Study

To better demonstrate the effectiveness of MCSEO, we train a BERT-based variant using only l_i^{text} and $l_i^{obj-phra}$ by setting α to zero in equation 7, using wiki + Flickr datasets. The results are presented in Table 2. Several observations can be drawn from the table. First, incorporating $l^{obj-phra}$ alongside the conventional textual objective l^{text} leads to a substantial improvement over the text-only baseline. Second, this two-loss variant performs comparably to combining l^{text} with the sentence-image-level objective $l^{img-cap}$ (which is equivalent to MCSE). This suggests that object-phrase-level

[1] https://github.com/IDEA-Research/Grounded-SAM-2

Method	Class.	Clust.	PairCl.	Retr.	STS	Avg.
MCSE (BERT)	65.30	36.63	76.12	26.04	74.48	51.57
MCSEO (BERT)	**67.13**	**37.43**	**77.10**	**27.23**	**75.72**	**52.79**
MCSE (RoBERTa)	68.25	35.15	77.88	25.77	74.86	51.99
MCSEO (RoBERTa)	**68.46**	**36.01**	**78.40**	**26.90**	**76.14**	**52.85**

Table 3: Results on 39 English MTEB tasks (wiki+Flickr), averaged within each task category. MCSEO improves over MCSE in all categories for both backbones.

alignment alone provides multimodal supervision of similar strength. Finally, the full MCSEO objective achieves the highest accuracy overall, showing that sentence-image-level and object-phrase-level alignments are complementary when used together.

To test whether the gains stem from correct alignment or merely from an additional loss term, we shuffle object-phrase correspondences within each image-caption pair. This variant not only loses the improvement but falls below MCSE, showing that the benefit requires correct alignment and that incorrect alignment is actively harmful. The reliability of this signal is supported by a manual evaluation of 1,000 extracted pairs, of which 96.7% are exactly correct (Appendix A.2).

4.4 Complementarity with KDMCSE

MCSEO and KDMCSE (Nguyen et al., 2024) address two different sources of noise in image–caption supervision: KDMCSE denoises *negatives* at the batch level through threshold filtering and an adaptive angular margin, whereas MCSEO refines *positives* within each pair through object–phrase alignment. The two are therefore orthogonal in principle, and can be combined by replacing the sentence–image term in Equation 7 with the full KDMCSE objective while retaining our object-phrase term:

$$l_i = l_i^{text} + \alpha\, l_i^{KDMCSE} + \beta\, l_i^{obj-phra}. \quad (8)$$

Table 4 reports the results on wiki+Flickr with the BERT backbone. Our reproduction of KDMCSE closely matches the originally reported score. MCSEO alone approaches KDMCSE without relying on a teacher model, indicating that fine-grained alignment is a strong supervision signal on its own. Combining the two outperforms either method individually, which empirically confirms that denoising negatives and refining positives are complementary rather than redundant. Implementation details are given in Appendix A.3.

Model	STS avg.
MCSE (reproduced)	$77.4_{\pm 0.6}$
MCSEO (ours)	$78.3_{\pm 0.5}$
KDMCSE (reproduced)	$78.7_{\pm 0.6}$
KDMCSE + MCSEO	$\mathbf{79.5}_{\pm 0.8}$

Table 4: Comparison and combination with KDMCSE on wiki+Flickr with the BERT backbone. Average Spearman's correlation over the seven STS tasks, reported as mean over five random seeds.

4.5 Evaluation Beyond STS

STS tasks measure only one aspect of sentence representation quality. To test whether the benefit of object–phrase alignment generalises, we additionally evaluate on 39 English tasks from MTEB v2 (Muennighoff et al., 2023), covering classification (8), clustering (9), pair classification (3), retrieval (10) and STS (9).

As shown in Table 3, MCSEO improves over MCSE on every one of the five task categories, for both backbones. The gains are not confined to the STS category on which the hyperparameters were tuned, suggesting that finer-grained multimodal supervision yields representations that transfer more broadly.

5 Conclusions

In this work, we propose MCSEO, a multimodal contrastive learning framework for sentence embeddings that introduces fine-grained supervision by aligning object-phrase pairs extracted from image-caption data. By leveraging existing detection and grounding models, MCSEO decomposes each image and caption into semantically grounded object-phrase pairs and applies an additional object-phrase-level contrastive loss during training. Extensive evaluations demonstrate that incorporating these finer-grained alignments significantly improves the quality of sentence embeddings across different backbone models.

Limitations

Despite the effectiveness of MCSEO, several limitations remain. Our experiments are conducted on image-caption datasets that are predominantly in English, and the generalizability of MCSEO to other languages or to more diverse visual domains (e.g., medical images, charts, or videos) remains to be verified. MCSEO depends on an external grounding and segmentation stack (Florence-2 and SAM2). Errors in this stack may propagate to the training signal. Such models are also likely biased toward common object categories; rare or abstract phrases are more often left ungrounded and therefore receive no fine-grained supervision, which may skew the benefit toward frequent concepts. Extracting objects and phrases introduces additional preprocessing overhead compared to standard image-caption contrastive learning, although this cost is incurred only once and does not affect inference.

Ethical Statements

This work uses publicly available image-caption datasets (e.g., Flickr30k) and pretrained models (SAM2, Florence2, BERT) strictly under their respective licenses and intended research use. We manually annotated 1,000 automatically extracted object–phrase pairs for quality evaluation; the annotation was performed by the authors and involved no external human subjects or newly collected data. Our method is designed to improve general-purpose sentence representation quality and does not target any sensitive or high-risk use case. We use LLM only for minor language polishing.

Acknowledgment

Kaiyan Zhao was supported by JST SPRING, Grant Number JPMJSP2108.

References

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)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Yiren Jian, Chongyang Gao, and Soroush Vosoughi. 2022. Non-linguistic supervision for contrastive learning of sentence embeddings. In *Advances in Neural Information Processing Systems*, volume 35, pages 35533–35548. Curran Associates, Inc.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pages 1760–1770. IEEE.

Juncheng Li, XIN HE, Longhui Wei, Long Qian, Linchao Zhu, Lingxi Xie, Yueting Zhuang, Qi Tian, and Siliang Tang. 2022a. Fine-grained semantically aligned vision-language pre-training. In *Advances in Neural Information Processing Systems*.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, and 1 others. 2022b. Grounded language-image pre-training. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10955–10965. IEEE.

Zhe Li, Laurence T. Yang, Xin Nie, Bocheng Ren, and Xianjun Deng. 2023. Enhancing sentence representation with visually-supervised multimodal pre-training. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 5686–5695, New York, NY, USA. Association for Computing Machinery.

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. Springer.

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. *Preprint*, arXiv:1907.11692.

Zhongtao Miao, Qiyu Wu, Kaiyan Zhao, Zilong Wu, and Yoshimasa Tsuruoka. 2024. Enhancing cross-lingual sentence embedding for low-resource languages with word alignment. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3225–3236, Mexico City, Mexico. Association for Computational Linguistics.

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, Dubrovnik, Croatia. Association for Computational Linguistics.

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, Mexico City, Mexico. Association for Computational Linguistics.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2641–2649.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, and Chaitanya Ryali et al. 2025. SAM 2: Segment anything in images and videos. In *The Thirteenth International Conference on Learning Representations*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding. *Preprint*, arXiv:1807.03748.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. 2024. Florence-2: Advancing a unified representation for a variety of vision tasks. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4818–4829.

Miaoran Zhang, Marius Mosbach, David Ifeoluwa Adelani, Michael A. Hedderich, and Dietrich Klakow. 2022. 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, Seattle, United States. Association for Computational Linguistics.

Kaiyan Zhao, Qiyu Wu, Xin-Qiang Cai, and Yoshimasa Tsuruoka. 2024. Leveraging multi-lingual positive instances in contrastive learning to improve sentence embedding. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 976–991, St. Julian's, Malta. Association for Computational Linguistics.

β	BERT	RoBERTa
— (MCSE)	$77.4_{\pm 0.6}$	$78.3_{\pm 0.4}$
0.001	$77.8_{\pm 0.4}$	$78.7_{\pm 0.6}$
0.005 (ours)	$\mathbf{78.3}_{\pm 0.5}$	$\mathbf{79.1}_{\pm 0.6}$
0.01	$78.1_{\pm 0.5}$	$78.9_{\pm 0.5}$
0.02	$77.7_{\pm 0.5}$	$78.9_{\pm 0.4}$
0.05	$77.6_{\pm 0.3}$	$78.5_{\pm 0.8}$

Table 5: Sensitivity of MCSEO to the object–phrase loss weight β (wiki+Flickr, $\alpha = 0.01$). Average Spearman's correlation over the seven STS tasks, mean over five seeds.

A Appendix

A.1 Experimental Setup

Specifically, we choose `bert-base-uncased` and `roberta-base` as the base models and use the `[CLS]` token as sentence embedding. To address the dimensional mismatch between textual embeddings and visual embeddings, we introduce an MLP layer that maps both modalities into a shared embedding space of 256 dimensions. We use cosine similarity as the similarity function and fix the temperature parameter to 0.05.

The detected objects are equipped with transparent backgrounds to avoid potential noise. For model training, we follow MCSE and set the weight of the image-caption contrastive loss α to the optimal value 0.01. The weight β for the object-phrase contrastive loss is selected via grid search from $\{0.001, 0.005, 0.01, 0.02, 0.05\}$, with $\beta = 0.005$ yielding the best results.

Table 5 reports the STS average as a function of the object–phrase loss weight β, on wiki+Flickr with $\alpha = 0.01$, averaged over five seeds. Every value of β in the searched range improves over MCSE for both backbones, indicating that the benefit does not hinge on a narrow choice of β, although the best value ($\beta = 0.005$) gives the largest gain.

To maximize the number of object-phrase pairs extracted from each example, we use the longest

Category	Proportion
Exact	96.7%
Partial / coarse	3.2%
Wrong	0.1%

Table 6: Manual evaluation of 1,000 randomly sampled object–phrase pairs from the training data. Only one pair out of 1,000 was judged wrong (one-sided 95% upper bound ≈0.5%).

caption provided for each image in Flickr30k. This strategy results in an average of 4.2 object-phrase pairs per image. During training, we exclude certain pairs from the object-phrase contrastive loss calculation if they are truncated due to length limitations or if they constitute the only one detected object-phrase pair in the image (insufficient for loss).

A.2 Annotation Protocol for Object-Phrase Pairs

Sampling. Annotation was carried out by the authors. We uniformly sample 1,000 object-phrase pairs from the pairs actually used for training on wiki+Flickr, i.e. after the exclusion rules described in Appendix A.1 have been applied. Each sampled instance is presented to the annotator as the source image with the segmentation mask overlaid, together with the full caption and the phrase span highlighted.

Labels. Each pair is assigned exactly one of three labels:

- **Exact.** The mask covers the referent of the phrase, and only that referent, up to minor boundary imprecision.
- **Partial / coarse.** The mask covers the correct referent, but imprecisely. For example, it captures only part of the object, includes a substantial amount of background, or covers only a subset of the instances denoted by a plural or collective phrase.
- **Wrong.** The mask covers a different entity from the one denoted by the phrase.

Results. The resulting distribution is reported in Table 6: 96.7% Exact, 3.2% Partial, and 0.1% Wrong (1 pair out of 1,000; one-sided 95% upper bound ≈0.5%). Strict precision is therefore 96.7% and lenient precision 99.9%.

Error analysis. The partial cases fall into two recurring patterns. First, plural or collective noun phrases (e.g. *a group of people*, *several bikes*) are frequently grounded to a single instance or to a subset of the instances they denote. Second, objects that are occluded or split across the image are over-segmented, so that the mask covers only the larger visible fragment. Neither pattern changes the identity of the referent, which is why we treat them as partial rather than wrong.

A.3 Implementation Details for the KDMCSE Comparison

For the comparison and combination reported in §4.4, we reproduce KDMCSE in our own codebase rather than quoting the published numbers, so that the two methods share the same data, batch size, evaluation protocol and seeds.

Teacher. Following Nguyen et al. (2024), we use a frozen CLIP ViT-B/32 as the teacher with pre-extracted features, replacing the ResNet50 encoder for this comparison only; segmented object crops are encoded by the same frozen visual encoder so that global and object-level representations share one space.

Hyperparameters. We adopt the values reported by Nguyen et al. (2024) for the KDMCSE-specific terms: angular margin $m = 0.125$ and similarity threshold 0.85. For the object–phrase term we keep our own $\beta = 0.005$, and α is set to $\alpha = 0.01$. All remaining settings are identical to those in Appendix A.1.

Sanity check. Our reproduced KDMCSE reaches an STS average of $78.7_{\pm 0.6}$ on wiki+Flickr with BERT, which closely matches the originally reported result, confirming that the reproduction is faithful.