A Cross-Model Alignment is not Sufficient for Uni-Model Retrieval task

Let $\mathcal{I}$ and $\mathcal{T}$ be the image and text domain, respectively. Let the image $I \in \mathcal{I}$ and the text $T \in \mathcal{T}$, forming an image-text pair. Additionally, let I' be another image from $\mathcal{I}$, constituting an image-image pair with I. Now we employ cross-model training, wherein the model learns the alignment solely between I and Y. Then, if we desire this model to predict the corresponding image I' by using I, it must first predict the associated text T and then employ T to predict I'. This is because the connection within the image data is solely established implicitly through the accompanying text. In this case, we treat T as a latent random variable.

This strategy is formalized by the following function:

$$f(I) \triangleq \mathbb{E}\left[\mathbb{E}\left[I'|T\right]\middle|I\right].$$

Furthermore, with the full knowledge of the dependency between I' and (I, T), we have an oracle predictor: $g(I, T) \triangleq \mathbb{E}\left[I'|I, T\right]$.

Inspired by the work of Tosh, Krishnamurthy, and Hsu (2021), we are now ready to give the following theorem, which directly leads to Proposition 1 in the main text.

Theorem 1. *Let* $\mathcal{E}_I = \mathbb{E}\left[\left(\mathbb{E}\left[I'|I\right] - g(I,T)\right)^2\right]$ *and* $\mathcal{E}_T = \mathbb{E}\left[\left(\mathbb{E}\left[I'|T\right] - g(I,T)\right)^2\right]$*, then*

$$\mathbb{E}\left[\left(f(I) - g(I,T)\right)^2\right] \leq \mathcal{E}_I + \mathcal{E}_T + \sqrt{\mathcal{E}_I \mathcal{E}_T},$$

This theorem suggest that the performance of $f(I)$ will be satisfactory when both I and T can independently predict I' effectively. On the one hand, this motivates us to train the uni-modal as this will reduce the term $\mathcal{E}_I$. On the other hand, by using the soft-label alignment technique, $\mathcal{E}_T$ could also be reduced. This is because in the cross-modal soft-label alignment, (I', T) is not simply treated as a negative image-text pair in the contrastive learning, which enables the model to better align T and I'.

Proof. For any $t > 0$, we have

$$\begin{aligned}
&\mathbb{E}\left[\left(f(I) - g(I,T)\right)^2\right] \\
=&\mathbb{E}\left[\left(\mathbb{E}\left[\mathbb{E}\left[I'|T\right]\middle|I\right] - \mathbb{E}\left[I'|I,T\right]\right)^2\right] \\
=&\mathbb{E}\left[\left(\mathbb{E}\left[\mathbb{E}\left[I'|T\right]\middle|I\right] - \mathbb{E}\left[I'|I\right] + \mathbb{E}\left[I'|I\right] - \mathbb{E}\left[I'|I,T\right]\right)^2\right] \\
\leq&(1+1/t)\underbrace{\mathbb{E}\left[\left(\mathbb{E}\left[\mathbb{E}\left[I'|T\right]\middle|I\right] - \mathbb{E}\left[I'|I\right]\right)^2\right]}_{A} \\
&+ (1+t)\underbrace{\mathbb{E}\left[\left(\mathbb{E}\left[I'|I\right] - \mathbb{E}\left[I'|I,T\right]\right)^2\right]}_{B}, \qquad (10)
\end{aligned}$$

where the last inequality is by the inequality of arithmetic and geometric means (i.e. AM–GM inequality).

Notice that $B = \mathcal{E}_I$. We then give an upper-bound for A:

$$\begin{aligned}
A =&\mathbb{E}\left[\left(\mathbb{E}\left[\mathbb{E}\left[I'|T\right]\middle|I\right] - \mathbb{E}\left[\mathbb{E}\left[I'|I,T\right]\right]\right)^2\right] \\
=&\mathbb{E}\left[\mathbb{E}^2\left[\mathbb{E}\left[I'|T\right] - \mathbb{E}\left[I'|I,T\right]\middle|I\right]\right] \\
\leq&\mathbb{E}\left[\left(\mathbb{E}\left[I'|T\right] - \mathbb{E}\left[I'|I,T\right]\right)^2\right] = \mathcal{E}_T,
\end{aligned}$$

where the inequality is obtained by applying Jensen's inequality to the square function.

Recall that Eq. (10) holds for any $t > 0$. Plugging A and B into Eq. (10), then

$$\begin{aligned}
\mathbb{E}\left[\left(f(I) - g(I,T)\right)^2\right] \leq& \inf_{t>0}(1+1/t)\mathcal{E}_T + (1+t)\mathcal{E}_I \\
=&\mathcal{E}_I + \mathcal{E}_T + \sqrt{\mathcal{E}_I \mathcal{E}_T},
\end{aligned}$$

where the last equality is obtained via optimizing over t.

This completes the proof. □

B Datasets and Settings

Datasets

image-text retrieval For image-text retrieval, we evaluate our approach on three datasets: Flickr30K (Young et al. 2014), MSCOCO (Lin et al. 2014), and ECCV Caption (Chun et al. 2022). Flickr30K consists of 31,000 images, each annotated with 5 sentences. Following the split in (Frome et al. 2013), we use 1,000 images for validation, 1,000 images for testing, and the rest for training. MSCOCO, on the other hand, contains 123,287 images, each with 5 annotated captions. It is divided into 113,287 training images, 5000 validation images, and 5000 test images. As for the ECCV Caption, it is a verified test set of MSCOCO with accurate annotations of false negatives and true negatives. It offers $\times 3.6$ image-to-text associations and $\times 8.5$ text-to-image associations compared to MSCOCO. We used the MSCOCO dataset as the training set for testing with the ECCV Caption dataset.

image retrieval For image retrieval experiments, our evaluation is conducted on the test sets of four widely used datasets: CUB (Welinder et al. 2010), SOP (Oh Song et al. 2016), In-Shop (Liu et al. 2016), and iNaturalist (Van Horn 2018). The number of examples and classes can be found in Table 6.

Dataset	Images	Classes
CUB	5,924	100
SOP	60,502	11,316
In-Shop	26,830	3,985
iNaturalist	136,093	2,452

Table 6: Dataset composition for evaluation in the image retrieval task.

semantic textual similarity In terms of semantic textual similarity, we evaluate our approach on seven STS tasks: STS 2012–2016(Agirre et al. 2012, 2013, 2014,

2015, 2016), STS Benchmark (Cer et al. 2017) and SICK-Relatedness (Marelli et al. 2014). These datasets provide labels between 0 and 5 indicating the semantic relatedness of sentence pairs.

Implementation Details

To validate the improved performance of our approach in cross-modal retrieval tasks, we executed a series of experiments involving three models: SGRAF (Diao et al. 2021), $\text{CLIP}_{\text{ViT-B/32, ViT-L/14@336}}$ (Radford et al. 2021), and $\text{X2VLM}_{\text{base}}$ (Zeng et al. 2023). The experimental parameters of the model are shown in Table 7.

SGRAF is an open-source, non-pretrained SOTA model for image-text retrieval. It consists of two models: SGR and SAF. To achieve uni-modal retrieval, global representations for images and texts are processed separately using two fully connected layers. For SAF, the hyperparameters used were $\alpha = 1.0$ and $\beta = 0.75$, while for SGR, they were set to $\alpha = 0.6$ and $\beta = 0.75$. In order to optimize training time, we implemented a faster version of SGRAF that utilized multiple GPUs while preserving the same model settings as the original work. For optimization, the AdamW optimizer was used with a weight decay of 1e-4. The learning rate was initially set to 2e-4 and decayed to 1e-5 following a cosine schedule. After training the model for 40 epochs on 4 V100 GPUs, the best-performing snapshot on the test set was selected for reporting the results.

On the other hand, CLIP is a pre-trained dual-encoder model that we experimented with using two different-sized models: ViT-B/32 and ViT-L/14@336. Both the text and image branches were enhanced with two FC layers, one for cross-modal retrieval and another for uni-modal retrieval. The hyperparameters for both models were set as $\alpha = 0.1$ and $\beta = 0.5$. Using the AdamW optimizer with a weight decay of 1e-4, the initial learning rate of 1e-5 decayed to 1e-6 according to the cosine annealing schedule. For the ViT-B/32 model, we fine-tuned it on 4 V100 GPUs for 30 epochs on the MSCOCO dataset and 40 epochs on the Flickr30K dataset. As for the ViT-L/14@336 model, fine-tuning was conducted on 4 A100 GPUs for 30 epochs on both datasets.

Lastly, X2VLM is an advanced pre-trained model with dual encoders and a fusion encoder. In our experiments, we focused on its base model. We made a modification to the image-text matching loss by incorporating soft-labels for hard negative samples while leaving the contrastive loss unchanged. Furthermore, an FC layer was added to both the image and text branches to bolster uni-modal retrieval. The hyperparameters for this experiment were $\alpha = 0.3$ and $\beta = 0.75$. During the fine-tuning process on 8 V100 GPUs, we adjusted the batch size and learning rate to $\frac{1}{4}$ of the original work while keeping other parameters unchanged. For the above three models, the temperature hyperparameters for uni-modal are initialized to 0.45 and learned together with the model.

C Detailed Experimental Results

Table 8 shows the complete results of our image-text retrieval experiments on the MSCOCO and Flickr30K datasets, while Table 9 shows the complete results of our image-text retrieval experiments on the ECCV Caption dataset. Table 10 and Table 11 respectively demonstrate the performance of our methods on the uni-modal tasks of image and text.

D Additional Computational Overhead

In Table 12, we report the resource consumption of each module in our method. The real-time feature extraction time consumption of the image uni-modal teacher model and that of the text uni-modal teacher model is 162.11 ms and 90.39 ms, respectively. It is worth noting that this step can be performed offline, where all features are pre-extracted and stored on the disk for loading, reducing the time consumption to 12.75 ms and 3.26 ms, respectively. Other modules (such as similarity calculation) are just element-wise operations of matrices with negligible time consumption (e.g., ≤ 0.3ms).

Furthermore, we also report the resource consumption of our methods using two different feature extraction strategies, namely real-time feature extraction and loading features from the disk. Notably, as shown in Table 13, when loading features from the disk, our method is only half a minute slower than the original method (i.e., "Contrastive loss" in Table 13) in one epoch. This minimal additional time overhead is almost negligible in practical model training. Note that when scaling up to a larger pre-training dataset, we suggest to use the real-time feature extraction, despite incurring more time overhead, as it provides more flexibility.

Model	#Params	epoch†[1]	epoch†[2]	GPU	τ‡	α	β
SGR	19M	40	40	4 V100	0.45	0.60	0.75
SAF	19M	40	40	4 V100	0.45	1.00	0.75
$\text{CLIP}_{\text{ViT-B/32}}$	152M	30	40	4 V100	0.45	0.10	0.50
$\text{CLIP}_{\text{ViT-L/14@336px}}$	429M	30	30	4 A100	0.45	0.10	0.50
X2VLM	253M	20	20	8 V100	0.45	0.30	0.75

Table 7: The parameters of the five models used for the experiment. †[1] represents the epoch of training with the MSCOCO dataset, †[2] represents the epoch of training with the Flickr30K dataset, and ‡ represents the initial parameter τ of our method.

Model	MSCOCO (5K Test Set)							Flickr30K (1K Test Set)						
	Image-to-Text			Text-to-Image			RSUM	Image-to-Text			Text-to-Image			RSUM
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
Faster-RCNN, ResNet-101, without pre-training														
SCAN	50.4	82.2	90.0	38.6	69.3	80.4	410.9	67.4	90.3	95.8	48.6	77.7	85.2	465.0
VSE∞	56.6	83.6	91.4	39.3	69.9	81.1	421.9	76.5	94.2	97.7	56.4	83.4	89.9	498.1
VSRN++	54.7	82.9	90.9	42.0	72.2	82.7	425.4	79.2	94.6	97.5	60.6	85.6	91.4	508.9
NAAF	58.9	85.2	92.0	42.5	70.9	81.4	430.9	**81.9**	**96.1**	98.3	61.0	85.3	90.6	513.2
SGR†[1]	57.3	83.2	90.6	40.5	69.6	80.3	421.5	76.6	93.7	96.6	56.1	80.9	87.0	490.9
SGR†[2]	56.0	83.6	91.2	39.8	69.1	79.8	419.6	74.2	93.4	97.0	56.1	82.2	88.3	491.1
+ **CSA**	56.6	84.4	91.6	40.3	70.9	81.3	425.1	78.7	95.2	97.7	57.1	84.0	89.4	502.0
+ **USA**	56.4	84.2	91.7	40.5	70.4	81.2	424.4	79.4	94.8	97.1	57.9	83.3	89.1	501.5
+ **CUSA**	57.4	84.5	92.0	40.9	71.2	81.9	427.9	79.3	94.9	97.5	58.4	84.2	89.5	503.7
SAF†[1]	55.5	83.8	91.8	40.1	69.7	80.4	421.3	75.6	92.7	96.9	56.5	82.0	88.4	492.1
SAF†[2]	56.1	83.4	91.2	39.7	69.7	80.7	420.9	74.6	94.8	97.0	55.8	81.9	88.2	492.4
+ **CSA**	56.2	84.8	92.4	40.3	71.3	82.0	426.9	77.3	94.9	97.9	58.1	84.6	90.1	502.9
+ **USA**	56.4	84.7	91.5	39.8	70.7	81.7	424.8	76.9	94.2	97.9	58.8	84.1	90.1	502.0
+ **CUSA**	55.6	84.7	92.3	40.8	71.7	82.4	427.5	77.8	95.0	98.0	58.5	83.9	90.3	503.5
SGRAF†[1]	58.8	84.8	92.1	41.6	70.9	81.5	429.7	78.4	94.6	97.5	58.2	83.0	89.1	500.8
SGRAF†[2]	58.7	85.0	91.6	41.5	70.6	81.0	428.3	76.4	94.5	97.5	58.4	83.8	89.4	500.0
+ **CSA**	59.8	86.3	**93.4**	42.9	73.0	83.3	438.5	80.7	95.9	98.4	**61.9**	**86.5**	91.1	**514.4**
+ **USA**	58.7	85.5	92.5	41.9	71.8	82.1	432.5	81.0	95.4	97.9	59.7	84.7	90.1	508.9
+ **CUSA**	**59.8**	**86.1**	93.3	**43.3**	**73.2**	**83.6**	**439.2**	81.4	95.6	**98.5**	61.0	86.1	**91.5**	514.1
Dual-Encoder, pre-training														
$\text{CLIP}_{\text{ViT-B/32}}$	56.3	81.7	89.4	42.8	71.2	81.1	422.6	78.7	95.4	98.0	66.3	88.6	93.1	520.0
+ **CSA**	**57.6**	82.2	89.7	44.1	72.2	81.9	427.7	80.4	94.6	97.8	**68.4**	**89.9**	**94.3**	525.4
+ **USA**	56.3	82.0	89.8	43.4	72.2	82.0	425.7	80.4	**95.8**	97.7	66.8	89.0	93.7	523.4
+ **CUSA**	57.3	**83.1**	**90.3**	**44.2**	**72.7**	**82.1**	**429.7**	**82.1**	95.3	**97.9**	67.5	89.6	93.9	**526.3**
$\text{CLIP}_{\text{ViT-L/14}}$‡	67.1	89.4	94.7	51.6	79.1	87.7	469.6	87.3	99.0	99.5	76.4	94.8	97.4	554.5
+ **CSA**	67.7	89.3	94.5	**52.7**	79.6	88.1	471.9	89.2	**99.2**	**99.8**	**78.0**	95.3	**97.9**	559.4
+ **USA**	66.6	89.8	**95.0**	51.5	79.2	87.9	470.0	88.5	98.9	99.5	75.8	94.5	97.6	554.8
+ **CUSA**	**67.9**	**90.3**	94.7	52.4	**79.8**	**88.1**	**473.1**	**90.8**	99.1	99.7	77.4	**95.5**	97.7	**560.2**
Dual Encoder + Fusion encoder reranking, pre-training														
$\text{BLIP}_{\text{base}}$	81.9	95.4	97.8	64.3	85.7	91.5	516.6	97.3	99.9	100.0	87.3	97.6	98.9	581.0
OmniVL	82.1	95.9	98.1	64.8	86.1	91.6	518.6	97.3	99.9	100.0	87.9	97.8	99.1	582.0
$\text{X2VLM}_{\text{base}}$	83.5	96.3	98.5	66.2	87.1	92.2	523.8	98.5	100.0	100.0	90.4	98.2	99.3	586.4
+ **CSA**	**83.6**	96.5	98.5	66.7	87.5	92.6	525.5	98.1	100.0	100.0	91.3	98.8	99.4	587.6
+ **USA**	82.8	96.5	**98.7**	66.7	87.6	92.7	525.0	97.9	100.0	100.0	90.6	98.7	99.5	586.7
+ **CUSA**	83.3	**96.6**	98.5	**67.1**	**87.6**	**92.7**	**525.8**	**98.5**	**100.0**	**100.0**	**91.3**	**98.8**	**99.5**	**588.1**
$\text{X2VLM}_{\text{large}}$	84.4	96.5	98.5	67.7	87.5	92.5	527.1	98.8	100.0	100.0	91.8	98.6	99.5	588.7

Table 8: Complete experimental results of image-text retrieval on MSCOCO and Flickr30K. †[1] denotes the improved results by the author compared to the original paper, †[2] denotes the results of reproducing the method, and ‡ represents the $\text{CLIP}_{\text{ViT-L/14@336px}}$ model.

Model	Image-to-Text			Text-to-Image		
	mAP@R	R-P	R@1	mAP@R	R-P	R@1
Faster-RCNN, ResNet-101, without pre-training						
SGR†[1]	26.8	38.7	70.3	42.2	51.2	83.6
+ **CSA**	27.9	39.8	73.0	43.4	52.4	83.2
+ **USA**	27.6	39.7	71.0	43.0	51.9	83.0
+ **CUSA**	28.1	40.0	72.4	44.0	53.0	83.4
SAF†[1]	26.6	38.5	69.6	43.1	52.0	83.8
+ **CSA**	27.7	39.8	72.3	44.4	53.3	84.9
+ **USA**	27.3	39.0	71.2	42.5	52.1	80.5
+ **CUSA**	27.4	39.8	71.4	44.4	53.6	84.6
SGRAF†[1]	28.1	39.8	72.3	43.7	52.5	84.4
+ **CSA**	29.5	41.1	**75.3**	46.1	54.6	**87.8**
+ **USA**	28.9	40.7	72.6	44.4	53.0	84.1
+ **CUSA**	**29.5**	**41.4**	74.5	**46.4**	**55.1**	85.7
Dual-Encoder, pre-training						
$\text{CLIP}_{\text{ViT-B/32}}$	28.5	39.4	72.5	41.7	50.8	83.0
+ **CSA**	29.5	40.5	**73.5**	44.6	53.2	85.7
+ **USA**	28.9	40.1	72.6	43.1	52.0	84.8
+ **CUSA**	**29.6**	**40.7**	72.0	**45.2**	**53.6**	**85.7**
$\text{CLIP}_{\text{ViT-L/14@336px}}$	32.8	43.4	79.7	45.5	54.2	87.2
+ **CSA**	33.6	44.1	80.0	47.2	55.7	88.1
+ **USA**	32.9	43.6	78.9	45.9	54.5	87.6
+ **CUSA**	**33.6**	**44.1**	**80.9**	**47.6**	**55.8**	**88.2**
Dual Encoder + Fusion encoder reranking, pre-training						
$\text{X2VLM}_{\text{base}}$ †[2]	36.6	45.2	89.7	43.8	51.2	93.5
+ **CSA**	37.6	46.4	89.5	47.7	55.3	94.1
+ **USA**	37.0	45.6	89.5	44.0	51.1	93.4
+ **CUSA**	**37.6**	**46.5**	**89.9**	**48.4**	**55.9**	**94.1**

Table 9: Complete experimental results of image-text retrieval on ECCV Caption. †[1] denotes the results of reproducing the method, and †[2] denotes the results from the checkpoint provided by the author.

Model	CUB	SOP	In-Shop	INaturalist	Avg.
Faster-RCNN, ResNet-101, without pre-training					
SGR†[1]	31.1	52.0	19.5	33.7	34.1
+ **CSA**	33.3	52.4	23.2	35.2	36.0
+ **USA**	**37.4**	**60.9**	**32.2**	**42.5**	**43.3**
+ **CUSA**	34.6	60.7	31.6	41.9	42.2
SAF†[1]	34.1	52.8	20.3	37.0	36.0
+ **CSA**	31.4	52.8	23.1	36.1	35.8
+ **USA**	38.7	**60.2**	**33.7**	44.1	**44.2**
+ **CUSA**	**39.9**	59.6	32.2	**44.6**	44.1
Dual-Encoder, pre-training					
$\text{CLIP}_{\text{ViT-B/32}}$	41.5	51.8	28.1	41.3	40.7
+ **CSA**	42.3	49.1	29.1	41.1	40.4
+ **USA**	**50.3**	**57.9**	**36.6**	**46.5**	**47.8**
+ **CUSA**	49.7	56.5	34.1	45.6	46.5
$\text{CLIP}_{\text{ViT-L/14@336px}}$	58.3	61.1	46.9	63.5	57.4
+ **CSA**	59.4	59.8	43.8	64.0	56.7
+ **USA**	**68.5**	**63.9**	**48.9**	**69.4**	**62.7**
+ **CUSA**	67.2	63.0	48.2	68.7	61.8
Dual Encoder + Fusion encoder reranking, pre-training					
$\text{X2VLM}_{\text{base}}$ †[2]	53.6	64.2	52.6	59.3	57.4
+ **CSA**	53.1	63.4	52.0	59.0	56.9
+ **USA**	**59.6**	**67.8**	**55.1**	**62.7**	**61.3**
+ **CUSA**	58.9	67.0	54.2	62.2	60.6
Uni-modal Alignment Using Data Augmentation					
$\text{TCL}_{\text{pretrain}}$ †[2]	27.7	73.1	52.1	43.8	49.2
$\text{TCL}_{\text{coco\_finetune}}$ †[2]	33.2	70.0	49.1	48.4	50.2
$\text{TCL}_{\text{flickr\_finetune}}$ †[2]	35.2	70.1	49.8	49.7	51.2

Table 10: Performance of image retrieval on 4 datasets. †[1] denotes the results of reproducing the method, and †[2] denotes the results from the checkpoint provided by the author.

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
Faster-RCNN, ResNet-101, without pre-training								
SGR†[1]	48.0	48.4	51.8	63.3	47.5	58.1	62.7	54.3
+ **CSA**	47.1	**49.0**	51.5	63.0	46.9	57.6	62.8	54.0
+ **USA**	**53.7**	42.7	51.2	**69.9**	**55.5**	**66.3**	64.8	57.7
+ **CUSA**	53.6	48.0	**53.8**	69.6	54.7	65.2	**64.9**	**58.5**
SAF†[1]	51.7	48.0	51.7	66.5	51.9	64.5	63.5	56.8
+ **CSA**	51.2	**51.5**	**52.5**	67.2	51.0	63.8	63.4	57.2
+ **USA**	**53.5**	44.3	51.7	69.9	**57.1**	**66.6**	**65.0**	**58.3**
+ **CUSA**	52.8	44.0	50.8	**70.2**	56.3	66.3	64.5	57.8
Dual-Encoder, pre-training								
$\text{CLIP}_{\text{ViT-B/32}}$	57.0	66.8	64.7	75.4	73.0	76.2	72.9	69.4
+ **CSA**	47.9	65.9	61.8	73.1	70.0	73.4	73.2	66.5
+ **USA**	**63.7**	71.8	68.6	**78.8**	**74.2**	**78.4**	**76.1**	73.1
+ **CUSA**	63.6	**72.6**	**69.3**	78.4	74.1	78.3	75.8	**73.2**
$\text{CLIP}_{\text{ViT-L/14@336px}}$	61.7	69.6	65.2	76.7	75.8	78.6	75.5	71.9
+ **CSA**	56.4	68.6	64.4	76.8	73.9	78.0	74.7	70.4
+ **USA**	**64.3**	**76.1**	**69.5**	**81.9**	**76.1**	**80.1**	**75.5**	**74.8**
+ **CUSA**	64.1	75.9	69.4	81.8	75.7	79.9	74.9	74.5
Dual Encoder + Fusion encoder reranking, pre-training								
$\text{X2VLM}_{\text{base}}$ †[2]	25.7	21.0	17.7	35.0	33.8	22.3	50.4	29.4
+ **CSA**	19.1	17.4	15.3	27.6	30.7	17.2	49.2	25.2
+ **USA**	**44.8**	**41.2**	**38.9**	**61.5**	**49.9**	**48.1**	75.5	**51.4**
+ **CUSA**	45.7	40.6	37.3	61.1	49.5	47.9	**76.2**	51.2
Uni-modal Alignment Using Data Augmentation								
$\text{TCL}_{\text{pretrain}}$ †[2]	25.8	38.5	29.8	26.9	46.6	42.3	67.2	39.6
$\text{TCL}_{\text{coco\_finetune}}$ †[2]	41.9	50.2	44.5	47.2	55.2	59.7	71.8	52.9
$\text{TCL}_{\text{flickr\_finetune}}$ †[2]	50.3	56.6	53.1	61.3	61.6	66.9	73.0	60.4

Table 11: Sentence embedding performance on STS tasks. †[1] denotes the results of reproducing the method and †[2] denotes the results from the checkpoint provided by the author.

Module	Time	Max GPU Mem	Model / File Disk Usage
Img-Fea extraction	162.11 ms	1831 MB	223 MB
Text-Fea extraction	90.39 ms	1037 MB	402 MB
Load Img-Fea from disk	12.75 ms	329 MB	15.9 GB
Load Text-Fea from disk	3.26 ms	311 MB	1.7 GB
Sims-Cal (I2I)	0.30 ms	459 MB	-
Sims-Cal (T2T)	0.28 ms	399 MB	-

Table 12: Resource consumption of each module. All results are the average of three runs with different random seeds. Each run has 1000 iterations with a batch size of 128.

Method	Time	Max GPU Mem	Max RAM
Contrastive loss	12.13 min	13239 MB	48 GB
Image Feature: Disk + Text Feature: Disk			
+ CSA	12.53 min	13362 MB	
+ USA	12.62 min	13369 MB	116 GB
+ CUSA	12.68 min	13371 MB	
Image Feature: Disk + Text Feature: Real-Time			
+ CSA	14.17 min	14233 MB	
+ USA	14.25 min	14145 MB	103 GB
+ CUSA	14.30 min	14146 MB	
Image Feature: Real-Time + Text Feature: Disk			
+ CSA	15.67 min	14296 MB	
+ USA	15.77 min	14308 MB	71 GB
+ CUSA	15.83 min	14310 MB	
Image Feature: Real-Time + Text Feature: Real-Time			
+ CSA	16.77 min	15127 MB	
+ USA	16.87 min	15138 MB	67 GB
+ CUSA	16.90 min	15139 MB	

Table 13: Resource consumption of training one epoch using various methods with $\text{CLIP}_{\text{ViT-B/32}}$. All results represent the average of three runs. Each run involves training models on the MSCOCO dataset using 4 V100 GPUs, a batch size of 128, and 5 epochs.

References

Agirre, E.; Banea, C.; Cardie, C.; Cer, D.; Diab, M.; Gonzalez-Agirre, A.; Guo, W.; Lopez-Gazpio, I.; Maritxalar, M.; Mihalcea, R.; et al. 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)*, 252–263.

Agirre, E.; Banea, C.; Cardie, C.; Cer, D.; Diab, M.; Gonzalez-Agirre, A.; Guo, W.; Mihalcea, R.; Rigau, G.; and Wiebe, J. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, 81–91.

Agirre, E.; Banea, C.; Cer, D.; Diab, M.; Gonzalez Agirre, A.; Mihalcea, R.; Rigau Claramunt, G.; and Wiebe, J. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval-2016. 10th International Workshop on Semantic Evaluation; 2016 Jun 16-17; San Diego, CA. Stroudsburg (PA): ACL; 2016. p. 497-511.* ACL (Association for Computational Linguistics).

Agirre, E.; Cer, D.; Diab, M.; and Gonzalez-Agirre, A. 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)*, 385–393.

Agirre, E.; Cer, D.; Diab, M.; Gonzalez-Agirre, A.; and Guo, W. 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*, 32–43.

An, X.; Deng, J.; Yang, K.; Li, J.; Feng, Z.; Guo, J.; Yang, J.; and Liu, T. 2023. Unicom: Universal and Compact Representation Learning for Image Retrieval. In *ICLR*.

Cer, D.; Diab, M.; Agirre, E.; Lopez-Gazpio, I.; and Specia, L. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Chen, J.; Hu, H.; Wu, H.; Jiang, Y.; and Wang, C. 2021. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15789–15798.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 1597–1607. PMLR.

Chun, S.; Kim, W.; Park, S.; Chang, M.; and Oh, S. J. 2022. Eccv caption: Correcting false negatives by collecting machine-and-human-verified image-caption associations for ms-coco. In *European Conference on Computer Vision*, 1–19. Springer.

Chun, S.; Oh, S. J.; de Rezende, R. S.; Kalantidis, Y.; and Larlus, D. 2021. Probabilistic Embeddings for Cross-Modal Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8415–8424.

Conneau, A.; and Kiela, D. 2018. SentEval: An Evaluation Toolkit for Universal Sentence Representations. arXiv:1803.05449.

Diao, H.; Zhang, Y.; Ma, L.; and Lu, H. 2021. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 1218–1226.

Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.; and Mikolov, T. 2013. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26.

Gao, T.; Yao, X.; and Chen, D. 2022. SimCSE: Simple Contrastive Learning of Sentence Embeddings. arXiv:2104.08821.

Gao, Y.; Liu, J.; Xu, Z.; Wu, T.; Liu, W.; Yang, J.; Li, K.; and Sun, X. 2023. SoftCLIP: Softer Cross-modal Alignment Makes CLIP Stronger. arXiv:2303.17561.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.

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

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 12888–12900. PMLR.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705.

Li, K.; Zhang, Y.; Li, K.; Li, Y.; and Fu, Y. 2022b. Image-text embedding learning via visual and textual semantic reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1): 641–656.

Li, Z.; Guo, C.; Feng, Z.; Hwang, J.-N.; and Du, Z. 2023a. Integrating Language Guidance into Image-Text Matching for Correcting False Negatives. *IEEE Transactions on Multimedia*, 1–14.

Li, Z.; Guo, C.; Feng, Z.; Hwang, J.-N.; Jin, Y.; and Zhang, Y. 2022c. Image-Text Retrieval with Binary and Continuous Label Supervision. arXiv:2210.11319.

Li, Z.; Guo, C.; Wang, X.; Feng, Z.; and Wang, Y. 2023b. Integrating Listwise Ranking into Pairwise-based Image-Text Retrieval. arXiv:2305.16566.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 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*, 740–755. Springer.

Liu, Z.; Luo, P.; Qiu, S.; Wang, X.; and Tang, X. 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1096–1104.

Marelli, M.; Bentivogli, L.; Baroni, M.; Bernardi, R.; Menini, S.; and Zamparelli, R. 2014. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, 1–8.

Oh Song, H.; Xiang, Y.; Jegelka, S.; and Savarese, S. 2016. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4004–4012.

Parekh, Z.; Baldridge, J.; Cer, D.; Waters, A.; and Yang, Y. 2021. Crisscrossed Captions: Extended Intramodal and Intermodal Semantic Similarity Judgments for MS-COCO. arXiv:2004.15020.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv:1908.10084.

Song, K.; Tan, X.; Qin, T.; Lu, J.; and Liu, T.-Y. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33: 16857–16867.

Tosh, C.; Krishnamurthy, A.; and Hsu, D. 2021. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, 1179–1206. PMLR.

van den Oord, A.; Li, Y.; and Vinyals, O. 2019. Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748.

Van Horn, G. 2018. Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, volume 2, 5.

Wang, J.; Chen, D.; Wu, Z.; Luo, C.; Zhou, L.; Zhao, Y.; Xie, Y.; Liu, C.; Jiang, Y.-G.; and Yuan, L. 2022. Omnivl: One foundation model for image-language and video-language tasks. *Advances in neural information processing systems*, 35: 5696–5710.

Welinder, P.; Branson, S.; Mita, T.; Wah, C.; Schroff, F.; Belongie, S.; and Perona, P. 2010. Caltech-UCSD birds 200.

Yang, J.; Duan, J.; Tran, S.; Xu, Y.; Chanda, S.; Chen, L.; Zeng, B.; Chilimbi, T.; and Huang, J. 2022. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15671–15680.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 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.

Zeng, Y.; Zhang, X.; Li, H.; Wang, J.; Zhang, J.; and Zhou, W. 2023. X^2-VLM: All-In-One Pre-trained Model For Vision-Language Tasks. arXiv:2211.12402.

Zhang, K.; Mao, Z.; Wang, Q.; and Zhang, Y. 2022. Negative-aware attention framework for image-text matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15661–15670.