MCDISP-ALIGN: MULTI-CAPTION SEMANTIC DISPERSION-GUIDED DISTRIBUTION ALIGNMENT FOR ONE-TO-MANY VISION-LANGUAGE REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language representation learning supports multimodal applications by aligning visual and textual semantics, but existing methods face two challenges. Deterministic methods embed samples as single points; aligning semantically distinct captions to a single image point may compress their differences and cause semantic distortion. Conversely, while probabilistic methods model samples as distributions, directly aligning an image distribution with individual caption distributions that primarily contain semantically similar points can restrict the image's semantic range, leading to semantic collapse. To address these challenges, we propose MCDisp-Align, a probabilistic framework that leverages multi-caption semantic dispersion to guide image distribution learning. MCDisp-Align constructs a caption-set distribution that captures the shared semantic center, semantic range, and dominant directions of variation across captions. By combining set-level distribution alignment with caption-center coverage, MCDisp-Align aims to reduce semantic distortion by preserving meaning differences between captions, and to reduce semantic collapse by guiding the image distribution to cover the meanings expressed by those captions. On MSCOCO and Flickr30K, MCDisp-Align achieves the best or tied-best one-to-many retrieval scores among all compared methods and improves all six standard retrieval scores over fine-tuned CLIP on MSCOCO. Geometric analysis further shows that caption representations retain variation within each image while remaining well separated across images, supporting the utility of caption-set supervision. Code will be released upon acceptance.

1 INTRODUCTION

Recent advances in large-scale vision-language models have substantially improved multimodal representation learning, enabling visual and textual information to be aligned in a shared semantic space (Li et al., 2021). This progress has supported applications including image-text retrieval (Lee et al., 2018), multimodal understanding (Kim et al., 2021), and other cross-modal tasks. The effectiveness of these applications depends on the quality of semantic alignment: related visual and textual content should be brought together, while meaningful semantic differences should be preserved and unrelated content should remain well separated.

Vision-language representation methods typically use deterministic representations, which map each image and caption to a point in an embedding space and align them via point-to-point matching (Radford et al., 2021). However, as illustrated in Figure 1(a), these captions share the main visual content but differ in their semantic focus. Figure 1(b) shows that aligning these captions with a single image point may compress their semantic distinctions, potentially leading to *semantic distortion* in the representation space. In practice, such distortion could reduce sensitivity to fine-grained information: retrieval systems may miss captions that emphasize specific attributes or relations, and multimodal question answering systems may overlook visual evidence needed to distinguish similar answers.

Probabilistic representation methods offer a more flexible alternative by modeling images and captions as distributions (Chun et al., 2025; Venkataramanan et al., 2025). However, distributional modeling

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

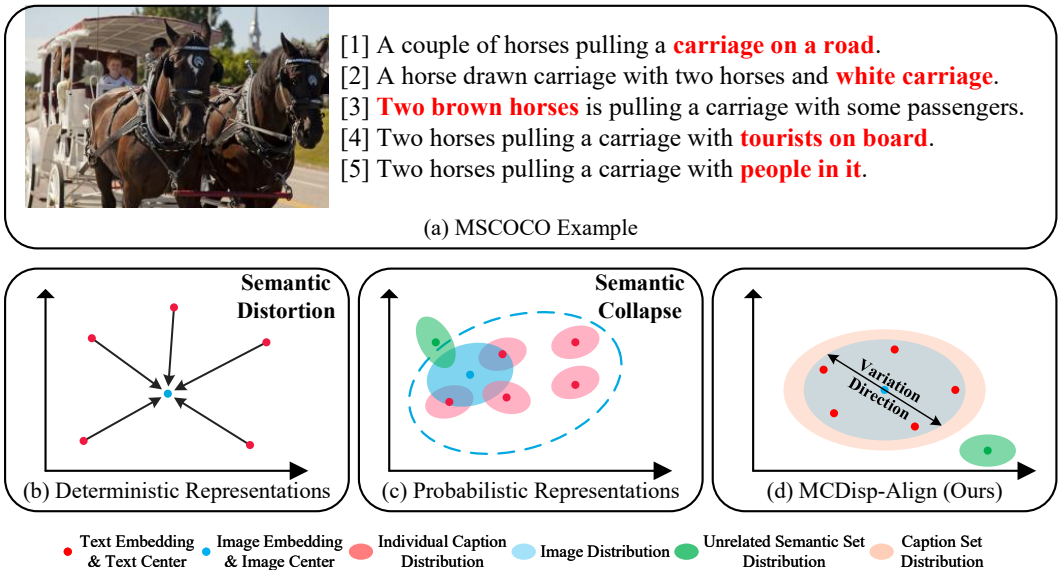


Figure 1: **Motivation for MCDisp-Align**. Comparison of (b) deterministic representations, (c) probabilistic representations, and (d) MCDisp-Align on the multi-caption example in (a).

alone may not fully capture the semantic differences among an image's captions. An image contains rich semantic content, including objects, attributes, relations, and background details, whereas an individual caption typically describes only part of this content. The image distribution is intended to accommodate multiple semantic aspects, whereas an individual caption distribution primarily contains points with meanings similar to that of the caption. As shown in Figure 1(c), directly aligning these distributions without accounting for this difference can restrict the image distribution's semantic range and suppress information beyond an individual caption's scope. We call this loss of semantic coverage in the image representation *semantic collapse*. This motivates using multiple complementary captions to guide the image distribution to cover their distinct semantic aspects.

Semantic distortion and semantic collapse highlight a central challenge for vision-language alignment: preserving the semantic diversity across an image's captions while maintaining separation from other images. To address this issue, we observe that multiple captions for the same image provide structural guidance. While individual captions emphasize different aspects, together they capture three key properties: a shared *semantic center*, a *semantic range* of complementary meanings, and *dominant directions of variation* that reflect how these meanings differ.

Based on this observation, we propose **MCDisp-Align**, a probabilistic representation learning framework that uses the caption-set structure to guide image distribution learning. Specifically, MCDisp-Align constructs a caption-set distribution that captures the shared semantic center, semantic range, and dominant directions of variation across captions. As illustrated in Figure 1(d), MCDisp-Align allows the image distribution to cover the different meanings expressed by multiple captions without directly pulling their text representations toward the same point. This helps preserve semantic differences across captions and reduce *semantic distortion*. The caption set further provides its semantic range and variation directions to guide image distribution learning, which helps mitigate *semantic collapse* while keeping different images well separated.

MCDisp-Align consists of three modules. First, the **probabilistic representation module** maps each image and caption to a Gaussian distribution, so that the image representation can capture a semantic range rather than a single point. Second, the **caption-set distribution construction module** merges the caption distributions of an image into one Gaussian through moment matching. The merged Gaussian captures both the overall center and the spread of the captions, providing a supervision target for the image distribution. Third, the **distribution matching and alignment module** aligns the image distribution with this target and encourages coverage of each individual caption center, while bidirectional contrastive matching keeps distinct images and caption sets well separated. Since the caption-derived targets are detached, these alignment constraints guide the image distribution without directly contracting the caption space, thereby preserving the diversity of caption representations.

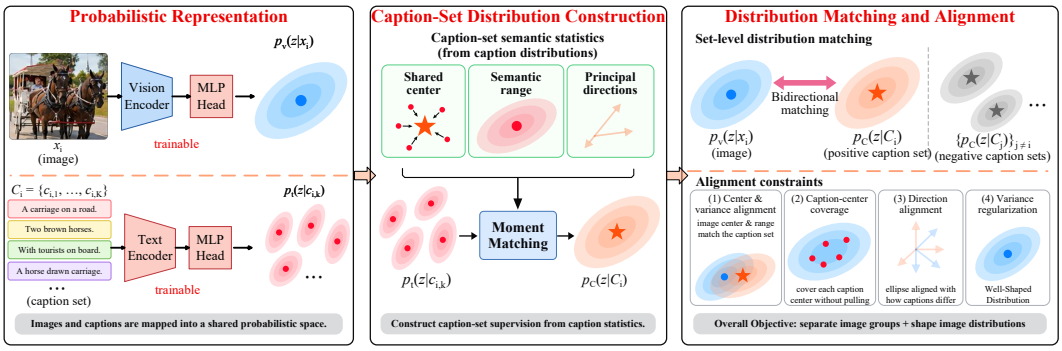


Figure 2: Overview of MCDisp-Align.

To evaluate MCDisp-Align, we focus on image-text retrieval, where each image is annotated with multiple captions and the task naturally exhibits a one-to-many cross-modal setting, using the benchmark datasets MSCOCO (Lin et al., 2014) and Flickr30K (Young et al., 2014). MCDisp-Align achieves the best or tied-best one-to-many retrieval scores on both benchmarks and improves over fine-tuned CLIP on all reported standard retrieval metrics on MSCOCO.

Our contributions are summarized as follows:

- We highlight two potential limitations in existing vision-language representations: semantic distortion in deterministic methods and semantic collapse in probabilistic methods.
- We propose MCDisp-Align, which leverages multi-caption semantic dispersion to supervise image distribution learning, combining caption-level coverage with set-level alignment.
- We evaluate retrieval and caption geometry on MSCOCO and Flickr30K, demonstrating one-to-many matching and standard-retrieval gains over fine-tuned CLIP on MSCOCO.

2 Method: Multi-Caption Semantic Dispersion-Guided Distribution Alignment

To mitigate semantic distortion and semantic collapse, MCDisp-Align uses the semantic variation across an image's captions to guide image distribution learning. As shown in Figure 2, MCDisp-Align consists of three modules: (1) **Probabilistic Representation** maps images and individual captions to Gaussian distributions, whose means represent semantic centers and covariances model variation around them; (2) **Caption-Set Distribution Construction** aggregates all caption distributions of an image via moment matching, summarizing their shared center, per-dimension spread, and principal directions of variation; and (3) **Distribution Matching and Alignment** distinguishes matching image-caption groups while using these statistics to supervise the image mean, diagonal variance, and low-rank covariance. A caption-level coverage constraint further encourages the image distribution to include the centers of its individual captions.

2.1 Probabilistic Representation

Each caption has a particular descriptive focus, while the image representation needs to accommodate differences across captions. We use diagonal covariances to model variation around individual caption centers. For images, an additional low-rank component models correlations across dimensions, allowing the distribution to extend along directions in which its captions differ.

Formally, for each image x_i and its associated caption set $\mathcal{C}_i = \{c_{i,k}\}_{k=1}^{K}$, we extract visual and textual features using encoders. Lightweight projection heads then parameterize the corresponding Gaussian distributions in a shared D-dimensional space:

$$p_v(\mathbf{z} \mid x_i) = \mathcal{N}(\boldsymbol{\mu}_i^v, \boldsymbol{\Sigma}_i^v), \qquad \boldsymbol{\Sigma}_i^v = \mathrm{Diag}(\mathbf{v}_i^v) + \mathbf{U}_i \mathbf{U}_i^\top,$$
$$p_t(\mathbf{z} \mid c_{i,k}) = \mathcal{N}(\boldsymbol{\mu}_{i,k}^t, \mathrm{Diag}(\mathbf{v}_{i,k}^t)), \tag{1}$$

where $\boldsymbol{\mu}_i^v, \boldsymbol{\mu}_{i,k}^t \in \mathbb{R}^D$ denote the semantic centers of the image and the k-th caption, respectively. $\mathbf{v}_i^v, \mathbf{v}_{i,k}^t \in (0,\infty)^D$ denote the positive variance vectors, and $\boldsymbol{\Sigma}_{i,k}^t = \mathrm{Diag}(\mathbf{v}_{i,k}^t)$ denotes the caption covariance. For the image distribution, $\mathbf{U}_i \in \mathbb{R}^{D\times r}$ is a low-rank factor with $r \ll D$.

For an individual caption, the covariance describes the range of points with meanings similar to that of the caption. Differences between captions are reflected in their centers. For images, $\mathbf{v}_i^v$ specifies the diagonal base component, while $\mathbf{U}_i\mathbf{U}_i^\top$ adds both marginal variance and cross-dimension correlations. This parameterization lets the image distribution cover different caption meanings while avoiding the cost of learning a full $D \times D$ covariance matrix.

Based on these individual caption distributions, we next construct a set-level distribution that summarizes the semantic variation across captions associated with the same image.

2.2 Caption-Set Distribution Construction

To guide an image distribution beyond an individual caption's content, we summarize the location and spread of the caption set. Averaging caption centers provides a shared center but loses information about their dispersion. We therefore form an equally weighted mixture of the K caption distributions and approximate it with a single Gaussian that preserves the mixture's mean and covariance:

$$p_{\mathcal{C}}(\mathbf{z} \mid \mathcal{C}_i) = \mathcal{N}(\boldsymbol{\mu}_i^{\mathcal{C}}, \boldsymbol{\Sigma}_i^{\mathcal{C}}).$$
$$\boldsymbol{\mu}_i^{\mathcal{C}} = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{\mu}_{i,k}^t, \qquad \boldsymbol{\Sigma}_i^{\mathcal{C}} = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{\Sigma}_{i,k}^t + \mathrm{Cov}_k(\boldsymbol{\mu}_{i,k}^t), \tag{2}$$

where Cov_k denotes the covariance of the caption centers around their set mean, computed with normalization $1/K$. The mean $\boldsymbol{\mu}_i^{\mathcal{C}}$ is the average of the caption centers, while the covariance $\boldsymbol{\Sigma}_i^{\mathcal{C}}$ combines two sources of semantic variation. The first term averages the learned covariances of the individual caption distributions, reflecting the spread of similar meanings around each caption center. The second term measures the dispersion of caption centers around their shared mean, capturing differences in what the captions describe, such as the horses' color or the passengers in a carriage. Together, these terms summarize the range of meanings expressed by the caption set and guide the image covariance to accommodate this range, helping mitigate semantic collapse.

These statistics shape the image distribution: the set mean provides a target for its center, the marginal variances control its base spread, and differences between caption centers determine the directions of its low-rank component. While this guidance shapes each image distribution individually, training must also distinguish matching image-caption-set pairs from mismatched ones. The next module combines caption-guided alignment with contrastive matching to achieve both goals.

2.3 Distribution Matching and Alignment

While the caption-set statistics define how an image distribution should accommodate its captions, training must keep different images well separated. We therefore optimize MCDisp-Align using two complementary objectives: a set-level matching loss $\mathcal{L}_{\mathrm{match}}$ to separate distinct image-caption groups across the batch, and an alignment loss $\mathcal{L}_{\mathrm{align}}$ to shape each image distribution using its caption set:

$$\mathcal{L} = \mathcal{L}_{\mathrm{match}} + \mathcal{L}_{\mathrm{align}}, \tag{3}$$

We describe the objective for one image and its caption set. We average caption-center coverage and caption variance regularization over the K captions; training averages the overall objective across the mini-batch. For image x_i in a mini-batch of B images, its own caption set is the positive target and the other $B-1$ sets are negatives. We score each image and caption-set pair using center cosine similarity scaled by their average variances:

$$s_{ij} = \frac{\cos(\boldsymbol{\mu}_i^v, \boldsymbol{\mu}_j^{\mathcal{C}})}{\tau\sqrt{1+\bar{v}_i^v}\sqrt{1+\bar{v}_j^{\mathcal{C}}}}, \tag{4}$$

where τ is a fixed temperature, and $\bar{v}_i^v$ and $\bar{v}_j^{\mathcal{C}}$ denote the average image base and caption-set variance, respectively. The centers indicate how well the image and caption set match. The variance terms

adjust the strength of this judgment: when either distribution has greater spread, the score reflects a weaker preference for or against the pair. We then apply bidirectional contrastive matching:

$$\mathcal{L}_{\text{match}} = -\frac{1}{2}[\log \frac{\exp(s_{ii})}{\sum_{j=1}^{B} \exp(s_{ij})} + \log \frac{\exp(s_{ii})}{\sum_{j=1}^{B} \exp(s_{ji})}], \tag{5}$$

which pulls matched images and caption sets together while pushing unmatched pairs apart.

$\mathcal{L}_{\text{align}}$ guides the image distribution to cover the different meanings expressed by these captions. It combines four terms. KL alignment guides the image center and base variances, while coverage encourages the image distribution to include individual caption centers. Subspace alignment guides the low-rank component toward directions in which captions differ, and variance regularization discourages excessively large or small image base and caption variances. We detach the caption targets in the first three terms using $\text{sg}(\cdot)$, so these constraints update only the image branch without directly changing the caption representations. Matching and variance regularization update both modalities. Omitting sample indices, we define:

$$\begin{aligned}\mathcal{L}_{\text{align}} = \frac{\alpha}{D} \text{KL}_{\text{diag}}(p_v \| p_{\mathcal{C}}) + \beta[\frac{d^2}{D} - 1]_+ + 2\gamma(r_0 - \|Q_v^\top Q_{\mathcal{C}}\|_F^2) \\ + \frac{\eta}{D}(\| \log \frac{\mathbf{v}^v}{v_0}\|_2^2 + \| \log \frac{\mathbf{v}^t}{v_0}\|_2^2),\end{aligned} \tag{6}$$

where α, β, γ, η weight the four terms, D is the embedding dimension.

The first term brings the image center and base variances closer to the caption-set mean and variances. We compute KL_{diag} from the diagonal image Gaussian with parameters $(\boldsymbol{\mu}^v, \mathbf{v}^v)$ to the diagonal caption-set target with parameters $(\boldsymbol{\mu}^{\mathcal{C}}, \text{diag}(\boldsymbol{\Sigma}^{\mathcal{C}}))$. We stop gradients through the caption-set targets and set a lower bound of 0.01 for each target variance for numerical stability.

The coverage term encourages the image distribution to include each caption center within the region $d^2 \leq D$. Here, d^2 is the squared Mahalanobis distance between the caption and image centers, computed using the full image covariance $\boldsymbol{\Sigma}^v + \epsilon \mathbf{I}$. The small positive constant ϵ improves numerical stability. A caption center outside this region adds a penalty that encourages the image distribution to move its center or adjust its spread. Once covered, the penalty becomes zero, and this term no longer encourages the two centers to move closer.

The third term helps the image's low-rank component follow the main directions in which caption centers differ. It compares the caption directions in $Q_{\mathcal{C}}$ with the image directions in Q_v, both of size $D \times r_0$. We form $Q_{\mathcal{C}}$ from the leading r_0 principal directions of the centered caption means. To obtain Q_v, we take the first r_0 columns of the orthonormal basis from the QR decomposition of $\mathbf{U}$, where $r_0 = \min(r, K-1, D)$. Maximizing $\|Q_v^\top Q_{\mathcal{C}}\|_F^2$ encourages the two subspaces to become more similar. This term does not match individual axes or set the variance along each direction.

The final term encourages the image base and caption variances to stay near $v_0 = 0.04$ by penalizing their differences from this value in log space. This discourages overly small or large variances. It does not directly limit the extra variance from $\mathbf{U}\mathbf{U}^\top$. Matching and variance regularization update the distribution heads and backbone encoders in both modalities throughout training. KL, coverage, and subspace alignment are scheduled in stages, and the matching loss is gradually allowed to update the variance predictions. The training-details appendix gives the full schedule.

For standard retrieval, caption sets are needed only during training. At inference, images and captions are encoded separately and matched via the cosine similarity between their distribution centers. For one-to-many evaluation, we aggregate the captions for each image into a single caption-set representation, so bidirectional matching reduces to a one-to-one comparison at the set level.

3 Experiments

We evaluate MCDisp-Align on one-to-many retrieval, matching images with representations of their caption sets. We then examine whether learning from caption-set distributions retains the ability to retrieve individual captions and images in standard retrieval. Caption-space analysis and ablations further assess the preservation of caption variation and the contribution of distribution alignment.

3.1 Experimental Setup

We use MSCOCO (Lin et al., 2014) and Flickr30K (Young et al., 2014), each with five reference captions per image. Our baselines include deterministic model CLIP (Radford et al., 2021) and probabilistic models ProLIP (Chun et al., 2025) and GroVE (Venkataramanan et al., 2025). CLIP and ProLIP are evaluated in zero-shot and fine-tuned settings; FT denotes fine-tuning on the target dataset. MCDisp-Align, CLIP, and GroVE use ViT-L/14, while ProLIP uses ViT-H/14 (Dosovitskiy et al., 2020). All methods use the same image pools, with text candidates defined by the retrieval setting.

We fine-tune the encoders and train the MCDisp-Align distribution heads with a batch size of 32 and a head learning rate of 5×10^{-5}. We stop training after three consecutive epochs without validation improvement and retain the best checkpoint. Images and captions are encoded independently at inference time. We report Recall@1/5/10 (R@1/5/10) for image-to-text (I2T) and text-to-image (T2I) retrieval as fractions, with differences expressed in percentage points.

3.2 One-to-Many Retrieval

Our main experiment evaluates one-to-many retrieval through bidirectional matching between images and caption sets. Given N images and their N caption sets, each set is treated as one item. The evaluation measures whether the correct image or caption set is retrieved. All methods use the same sets of five reference captions per image: I2T retrieves the corresponding caption set for an image, while T2I retrieves the corresponding image for a caption set. All methods match each image against an aggregated representation of its caption set: the probabilistic baselines ProLIP and GroVE aggregate caption distribution features; the deterministic baseline CLIP aggregates caption embedding features, and MCDisp-Align aggregates its caption distributions via moment matching.

Table 1: Bidirectional One-to-Many retrieval on MSCOCO and Flickr30K. Best results are in **bold**.

	I2T R@1	I2T R@5	I2T R@10	T2I R@1	T2I R@5	T2I R@10
			MSCOCO			
CLIP (zero-shot)	0.625	0.844	0.910	0.511	0.751	0.839
CLIP-FT	0.825	0.951	0.979	0.761	0.932	0.966
ProLIP (zero-shot)	0.630	0.840	0.902	0.524	0.748	0.833
ProLIP-FT	0.774	0.932	0.962	0.738	0.917	0.956
GroVE	0.573	0.819	0.893	0.444	0.713	0.805
MCDisp-Align	**0.852**	**0.974**	**0.992**	**0.839**	**0.970**	**0.988**
			Flickr30K			
CLIP (zero-shot)	0.799	0.958	0.987	0.731	0.930	0.975
CLIP-FT	0.900	0.986	0.994	0.884	0.984	**0.996**
ProLIP (zero-shot)	0.791	0.952	0.979	0.752	0.936	0.971
ProLIP-FT	0.857	0.977	0.994	0.845	0.974	0.991
GroVE	0.868	0.982	0.994	0.779	0.952	0.978
MCDisp-Align	**0.903**	**0.988**	**0.997**	**0.909**	**0.990**	**0.996**

Table 1 leads to the following observations:

(1) MCDisp-Align consistently improves bidirectional image-caption-set matching on both datasets. It achieves the best performance on eleven of the twelve reported evaluation metrics and ties with CLIP-FT on Flickr30K T2I R@10 (0.996). For I2T R@1, MCDisp-Align outperforms the strongest baseline, CLIP-FT, by 2.7 points in MSCOCO and 0.3 points in Flickr30K. Since all methods are evaluated using the same caption sets, these results show that MCDisp-Align can better learn the relationships between an image and its multiple captions.

(2) The improvement is especially clear when the goal is to rank the correct image first given a caption set. T2I R@1 reaches 0.839 on MSCOCO and 0.909 on Flickr30K, which is 7.8 and 2.5 points higher than CLIP-FT, respectively. These improvements are larger than those for I2T in both datasets, which shows that MCDisp-Align is particularly helpful when using several related captions to find the correct image. This result also matches our goal of learning an image distribution that more effectively captures the meaning shared across its caption set.

Together, these results show that MCDisp-Align performs well for bidirectional image-caption-set matching. Next, we test whether this improvement also applies to standard retrieval and study the effect of $\mathcal{L}_{\text{align}}$ through controlled ablation in Section 3.5.

3.3 Standard Image-Text Retrieval

Our main focus is learning one-to-many image-text relationships through caption-set supervision. As a supplementary evaluation, we examine whether the learned representations retain the ability to match individual images and captions in standard retrieval. At inference time, we rank candidate matches by the cosine similarity between image and caption centers, without aggregating captions.

We report this supplementary comparison on MSCOCO. Table 2 leads to the following observations:

Table 2: Standard image-text retrieval on MSCOCO. Best results are in **bold**.

Model	I2T R@1	I2T R@5	I2T R@10	T2I R@1	T2I R@5	T2I R@10
CLIP (zero-shot)	0.560	0.792	0.859	0.359	0.602	0.705
CLIP-FT	0.676	0.881	0.940	0.546	0.803	0.880
ProLIP (zero-shot)	0.564	0.786	0.858	0.360	0.591	0.687
ProLIP-FT	**0.731**	**0.911**	0.954	**0.565**	0.805	0.879
GroVE	0.514	0.756	0.838	0.305	0.559	0.670
MCDisp-Align	0.696	0.906	**0.955**	0.553	**0.815**	**0.891**

(1) MCDisp-Align improves individual image-caption matching while using the same encoder architecture and pretrained CLIP ViT-L/14 backbone as CLIP. It outperforms both zero-shot CLIP and CLIP-FT across all six metrics. Compared with CLIP-FT, I2T R@1 increases from 0.676 to 0.696 and T2I R@1 from 0.546 to 0.553. This comparison supports the benefit of the proposed approach without relying on a larger encoder or a different pretrained backbone, and shows that caption-set supervision can also improve standard retrieval.

(2) ProLIP-FT performs best at rank one, achieving I2T and T2I R@1 of 0.731 and 0.565, respectively. However, it uses a different pretrained ViT-H/14 backbone, whose greater capacity and pretraining may contribute to this advantage. Therefore, this comparison does not isolate the effect of the alignment method. MCDisp-Align remains competitive, outperforming ProLIP-FT on I2T R@10 and T2I R@5/10, while its consistent gains over CLIP provide a more direct comparison with the same encoder and pretrained backbone under otherwise comparable evaluation settings.

These results show that MCDisp-Align performs well in one-to-many retrieval while maintaining strong performance in standard image–text retrieval.

3.4 Further Analysis

We examine how much caption representations vary within each image and how well they separate different images in the representation space. We use the same image-caption samples for all models, comparing the text embeddings of CLIP and GroVE with the means of the individual-caption distributions produced by ProLIP and MCDisp-Align.

Within-Set Dispersion (WSD) is the mean cosine distance between caption representations from the same image, while Between-Set Dispersion (BSD) measures this distance across different images. Relative Set Dispersion (RSD), defined as $\frac{\text{WSD}/\text{BSD}+(\text{BSD}-\text{WSD})}{2}$. RSD averages the ratio and difference terms to summarize the spread of caption representations within each set and their separation across sets. The ratio term becomes small when WSD is low relative to BSD, while the difference term measures how much BSD exceeds WSD. Together, these terms provide a geometric description of how caption representations vary within and across images.

Table 3 shows the following observations:

(1) For MSCOCO, MCDisp-Align has greater caption dispersion within each image alongside strong separation across images. Its WSD is 0.408, compared with 0.320 for CLIP-FT and 0.266 for ProLIP-FT, while its BSD reaches 0.996. Its RSD is higher than that of CLIP-FT (0.499 versus

Table 3: Mean cosine distances between caption representations within and across images.

	MSCOCO			Flickr30K		
	WSD	BSD	RSD	WSD	BSD	RSD
CLIP (zero-shot)	0.302	0.643	0.405	0.357	0.673	0.423
CLIP-FT	0.320	0.986	0.495	0.333	0.953	0.485
ProLIP (zero-shot)	0.332	0.801	0.442	**0.364**	0.766	0.439
ProLIP-FT	0.266	0.734	0.415	0.296	0.725	0.419
GroVE	0.209	0.488	0.354	0.253	0.510	0.377
MCDisp-Align	**0.408**	**0.996**	**0.499**	0.323	**0.985**	**0.495**

0.495), describing less compact caption sets relative to their separation. Together with the retrieval results, these findings suggest that MCDisp-Align preserves differences in meaning between captions of the same image while maintaining accurate image-text matching.

(2) For Flickr30K, MCDisp-Align retains within-set variation close to CLIP-FT (WSD of 0.323 versus 0.333), with greater between-set separation (BSD of 0.985 versus 0.953). Its RSD of 0.495 summarizes the relative compactness of this geometry. The retained caption variation is indicated by WSD remaining close to CLIP-FT, rather than by the lower ratio itself.

These comparisons show that caption representations can retain variation within each image while remaining well separated across images. The controlled ablation below tests whether adding the alignment loss reduces this variation in the learned representation space.

3.5 Ablation Study

We remove $\mathcal{L}_{\text{align}}$ but still fine-tune the encoders, keeping all other settings unchanged. We evaluate standard retrieval and caption-space geometry to test whether distribution alignment improves individual-caption matching without reducing caption variation.

Table 4: Retrieval performance of MCDisp-Align with and without $\mathcal{L}_{\text{align}}$. Best results are in **bold**.

Dataset	Model	I2T R@1	I2T R@5	I2T R@10	T2I R@1	T2I R@5	T2I R@10
MSCOCO	w/o $\mathcal{L}_{\text{align}}$	0.608	0.865	0.939	0.481	0.752	0.838
	MCDisp-Align	**0.696**	**0.906**	**0.955**	**0.553**	**0.815**	**0.891**
Flickr30K	w/o $\mathcal{L}_{\text{align}}$	0.691	0.904	0.964	0.597	0.861	0.924
	MCDisp-Align	**0.769**	**0.941**	**0.976**	**0.641**	**0.886**	**0.939**

Tables 4 and 5 show the following observations:

(1) With all other training settings unchanged, adding $\mathcal{L}_{\text{align}}$ improves standard retrieval on both datasets. The largest gains occur at R@1: I2T and T2I increase by 8.8 and 7.2 points on MSCOCO, and by 7.8 and 4.4 points on Flickr30K. R@5 and R@10 also improve in both directions, showing that the alignment loss helps the model rank matching images and captions more accurately.

Table 5: Caption dispersion of MCDisp-Align with and without $\mathcal{L}_{\text{align}}$. Best results are in **bold**.

Dataset	Model	WSD	BSD	RSD
MSCOCO	w/o $\mathcal{L}_{\text{align}}$	0.402	**1.000**	**0.500**
	MCDisp-Align	**0.408**	0.996	0.499
Flickr30K	w/o $\mathcal{L}_{\text{align}}$	0.322	**0.985**	0.495
	MCDisp-Align	**0.323**	**0.985**	**0.495**

(2) Caption dispersion changes little as retrieval improves. WSD increases from 0.402 to 0.408 on MSCOCO and from 0.322 to 0.323 on Flickr30K, while BSD decreases slightly from 1.000 to 0.996 on MSCOCO and remains at 0.985 on Flickr30K. These results show that better retrieval does not require a reduction in average within-set variation, and that separation across images remains largely unchanged.

(3) This pattern is consistent with the roles of the two objectives. The matching loss separates different image-caption groups, while the KL, coverage, and direction-alignment terms use detached caption targets to shape the image representation. A caption center incurs no coverage penalty once it

falls inside the image distribution's coverage region. The model can therefore accommodate multiple valid captions without forcing their centers to coincide.

Removing $\mathcal{L}_{\text{align}}$ weakens retrieval on both datasets even though caption dispersion changes little, pointing to its contribution to cross-modal matching.

4 Related Work

4.1 Vision-language representation learning

Common vision-language architectures include fusion-based models and dual encoders. UNITER learns joint representations through modality fusion (Chen et al., 2020), while CLIP and ALIGN use separate encoders trained with large-scale contrastive learning (Radford et al., 2021; Jia et al., 2021; Lu et al., 2025). SigLIP replaces the softmax loss with a pairwise sigmoid loss (Zhai et al., 2023), and SigLIP 2 extends it to multilingual and fine-grained settings (Tschannen et al., 2025).

Beyond global alignment, local alignment methods combine image and text tokens (Asokan et al., 2025), use text-conditioned pooling (Xiao et al., 2025), or train with region-text pairs and hard negatives (Xie et al., 2025; 2026). GOAL, β-CLIP, and TIPSv2 match images and text at multiple levels of detail (Choi et al., 2025; Zohra et al., 2026; Cao et al., 2026). Other methods exploit visual structure (Ruan et al., 2026) or long captions (Lavoie et al., 2026), while FALCON reduces false negatives in one-to-many matching through adaptive negative mining (Kim et al., 2026).

MCDisp-Align shares the dual-encoder design and the goal of preserving fine-grained semantics, focusing on variation among an image's valid captions. It uses caption-set dispersion to supervise the image distribution's semantic range while retaining independent encoding for standard retrieval.

4.2 Probabilistic vision-language representation learning

Probabilistic methods model ambiguity by mapping inputs to distributions. PCME introduces probabilistic embeddings (Chun et al., 2021), which PCME++ improves using a closed-form distance and better false-negative handling (Chun, 2024). Probabilistic representations have been expanded to composed retrieval (Tang et al., 2025), large-scale pre-training (Chun et al., 2025), probabilistic adapters (Upadhyay et al., 2023), and latent variable models (Venkataramanan et al., 2025). Other methods estimate uncertainty without learning new embeddings, using Bayesian projection layers (Baumann et al., 2026) or class-conditional Gaussian distributions (Lin et al., 2026). To account for normalized embeddings, AsymVLM models asymmetric uncertainty on the unit sphere (Ju et al., 2025), and REPVLM estimates uncertainty using density models (Ju et al., 2026).

MCDisp-Align uses variation across an image's captions to supervise its distribution, combining within-caption uncertainty and between-caption dispersion through moment matching. The resulting targets guide the image's semantic center, range, and dominant variation directions. Detaching these targets allows set-level alignment and caption-center coverage to shape the image distribution without directly pulling caption centers together.

5 Conclusion

In this paper, we present MCDisp-Align, which uses multi-caption semantic dispersion to guide image distribution learning for one-to-many vision-language alignment. Caption-set supervision helps reduce semantic distortion by shaping the image distribution without directly pulling semantically distinct caption centers together. Set-level alignment and caption-center coverage help mitigate semantic collapse by guiding the image distribution to cover the complementary meanings of its captions. On MSCOCO and Flickr30K, MCDisp-Align achieves the best or tied-best one-to-many retrieval scores among the methods, while retaining performance in standard image-text retrieval and outperforming fine-tuned CLIP on MSCOCO. Caption-space analysis and ablations show that these gains coexist with retained caption variation within images and separation across images.

AI USE STATEMENT

We used generative AI tools, including ChatGPT 6 Astra and GLM 5.3, to help translate texts, debug and optimize code, and find research information and relevant literature. These tools also helped us improve the paper's wording and readability and correct grammatical errors.

We did not use generative AI to develop research ideas or the conceptual framework, propose or refine hypotheses, formulate mathematical claims, or contribute to or write proofs. We also did not use it to design or provide feedback on the research methodology or experiments, generate synthetic datasets, clean or reformat datasets, conduct qualitative or thematic data analysis, or interpret experimental results.

We reviewed all AI-assisted work, checking the translations and language edits for accuracy and consistency with the intended scientific meaning. We verified the information and references against their sources and inspected and tested the code changes for correctness. We take responsibility for the final content of this work, including all AI-assisted text, claims, code, and other artifacts.

REFERENCES

Mothilal Asokan, Kebin Wu, and Fatima Albreiki. Finelip: Extending clip's reach via fine-grained alignment with longer text inputs. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 14495–14504, June 2025.

Anton Baumann, Rui Li, Marcus Klasson, Santeri Mentu, Shyamgopal Karthik, Zeynep Akata, Arno Solin, and Martin Trapp. Post-hoc probabilistic vision-language models. In *International Conference on Learning Representations (ICLR)*, 2026.

Bingyi Cao, Koert Chen, Kevis-Kokitsi Maninis, Kaifeng Chen, Arjun Karpur, Ye Xia, Sahil Dua, Tanmaya Dabral, Guangxing Han, Bohyung Han, Joshua Ainslie, Alex Bewley, Mithun Jacob, René Wagner, Washington Ramos, Krzysztof Choromanski, Mojtaba Seyedhosseini, Howard Zhou, and Andre Araujo. Tipsv2: Advancing vision-language pretraining with enhanced patch-text alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29325–29335, June 2026.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020.

Hyungyu Choi, Young Kyun Jang, and Chanho Eom. Goal: Global-local object alignment learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4070–4079, June 2025.

Sanghyuk Chun. Improved probabilistic image-text representations. In *International Conference on Learning Representations (ICLR)*, 2024.

Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Sanghyuk Chun, Wonjae Kim, Song Park, and Sangdoo Yun. Probabilistic language-image pre-training. In *International Conference on Learning Representations (ICLR)*, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 2021. URL http://proceedings.mlr.press/v139/jia21b.html.

Li Ju, Mayank Nautiyal, Andreas Hellander, Ekta Vats, and Prashant Singh. Epistemic uncertainty quantification for pre-trained VLMs via riemannian flow matching. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=E8Q093jDug.

Li Ju et al. Exploiting the asymmetric uncertainty structure of pre-trained vlms on the unit hypersphere. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Myunsoo Kim, Seongwoong Shim, and Byung-Jun Lee. Falcon: False-negative aware learning of contrastive negatives in vision-language alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 701–711, June 2026.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:231839613.

Marc-Antoine Lavoie, Anas Mahmoud, Aldo Zaimi, Arsene Fansi Tchango, and Steven L. Waslander. Clip is shortsighted: Paying attention beyond the first sentence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9524–9534, June 2026.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Neural Information Processing Systems*, 2021. URL https://api.semanticscholar.org/CorpusID:236034189.

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

Zhenxiang Lin, Maryam Haghighat, Will Browne, and Dimity Miller. Intra-class probabilistic embeddings for uncertainty estimation in vision-language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2327–2337, March 2026.

Jianglin Lu, Hailing Wang, Yi Xu, Yizhou Wang, Kuo Yang, and Yun Fu. Representation potentials of foundation models for multimodal alignment: A survey. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 16669–16684, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.843. URL https://aclanthology.org/2025.emnlp-main.843/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Zanxi Ruan, Songqun Gao, Qiuyu Kong, Yiming Wang, and Marco Cristani. Structxlip: Enhancing vision-language models with multimodal structural cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17292–17302, June 2026.

Haomiao Tang, Jinpeng Wang, Yuang Peng, Guanghao Meng, Ruisheng Luo, Bin Chen, Long Chen, Yaowei Wang, and Shu-Tao Xia. Modeling uncertainty in composed image retrieval via probabilistic embeddings. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1210–1222, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.61. URL https://aclanthology.org/2025.acl-long.61/.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

U. Upadhyay, S. Karthik, M. Mancini, and Z. Akata. Probvlm: Probabilistic adapter for frozen vision-language models. In *International Conference on Computer Vision (ICCV 2023)*, 2023.

Aishwarya Venkataramanan, Paul Bodesheim, and Joachim Denzler. Probabilistic embeddings for frozen vision-language models: Uncertainty quantification with gaussian process latent variable models. In *Proceedings of the Forty-First Conference on Uncertainty in Artificial Intelligence*, 2025.

Rui Xiao, Sanghwan Kim, Mariana-Iuliana Georgescu, Zeynep Akata, and Stephan Alaniz. Flair: Vlm with fine-grained language-informed image representations. In *CVPR*, 2025.

Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Gengshen Zhang, Dawei Leng, and Yuhui Yin. Fg-clip: Fine-grained visual and textual alignment. In *International Conference on Machine Learning*, pp. 68777–68793, 2025.

Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Ji Ao, Dawei Leng, and Yuhui Yin. Fg-clip 2: A bilingual fine-grained vision-language alignment model. In *International Conference on Machine Learning*, 2026.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, October 2023.

Fatimah Zohra, Chen Zhao, Hani Itani, and Bernard Ghanem. b-clip: Text-conditioned contrastive learning for multi-granular vision-language alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 680–689, June 2026.

A Hyperparameters and Training Details

We fix the temperature in Eq. 4 to $\tau = 0.07$ and set the base loss weights in Eq. 6 to $\alpha = 1.0$, $\beta = 0.5$, $\gamma = 0.01$, and $\eta = 0.01$. The low-rank dimension is set to $r = 4$, the inversion stabilizer to $\epsilon = 10^{-6}$, and the variance anchor to $v_0 = 0.04$. We use Adam with a weight decay of 10^{-4}. The backbone encoders are fine-tuned with a base learning rate of 10^{-6}, while the distribution heads use 5×10^{-5}. Both learning rates follow a cosine schedule with a one-epoch warmup. Each mini-batch contains 32 images, each paired with $K = 5$ captions.

Training runs for up to 10 epochs. Matching and variance regularization remain active throughout training and are the only active loss terms during the first two epochs. During these two epochs, we block gradients from the matching loss through the variance predictions. Over epochs 3–4, we linearly increase the KL loss weight from 0.05α toward α and the matching gradient scale for variance predictions from zero toward one. Both reach their full values at epoch 5, when the coverage term is also activated. Direction alignment is scheduled to begin at epoch 9 with a weight of 0.5γ, increasing to γ at epoch 10. We stop training after three consecutive epochs without validation improvement and retain the best checkpoint for evaluation.