MODEST-ALIGN: Data-Efficient Alignment for Vision-Language Models

Jiaxiang Liu[1,2] Yuan Wang[2] Jiawei Du[3,4]
Joey Tianyi Zhou[3,4] Mingkun Xu[*,1] Zuozhu Liu[*,2]

[1] Guangdong Institute of Intelligence Science and Technology, China

[2] ZJU-Angelalign R&D Center for Intelligence Healthcare, Zhejiang University, China

[3] Centre for Frontier AI Research (CFAR), Agency for Science, Technology and Research (A*STAR), Singapore

[4] Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore

{forworkliu}@gmail.com

Abstract

Cross-modal alignment aims to map heterogeneous modalities into a shared latent space, as exemplified by models like CLIP, which benefit from large-scale image-text pretraining for strong recognition capabilities. However, when operating in resource-constrained settings with limited or low-quality data, these models often suffer from overconfidence and degraded performance due to the prevalence of ambiguous or weakly correlated image-text pairs. Current contrastive learning approaches, which rely on single positive pairs, further exacerbate this issue by reinforcing overconfidence on uncertain samples. To address these challenges, we propose MODEST-ALIGN, a lightweight alignment framework designed for robustness and efficiency. Our approach leverages two complementary strategies—*Random Perturbation*, which introduces controlled noise to simulate uncertainty, and *Embedding Smoothing*, which calibrates similarity distributions in the embedding space. These mechanisms collectively reduce overconfidence and improve performance on noisy or weakly aligned samples. Extensive experiments across multiple benchmark datasets demonstrate that MODEST-ALIGN outperforms state-of-the-art methods in retrieval tasks, achieving competitive results with over 100× less training data and 600× less GPU time than CLIP. Our method offers a practical and scalable solution for cross-modal alignment in real-world, low-resource scenarios.

1 Introduction

Multimodal learning, by integrating different types of data modalities, enhances a model's perception and understanding capabilities, facilitating cross-modal information interaction and integration (Radford et al., 2021; Vouitsis et al., 2024; Zhai et al., 2022; Liu et al., 2023b,a; Yang et al., 2021; Girdhar et al., 2023; Liu et al., 2024a,b). Recent advancements in multimodal machine learning have shown
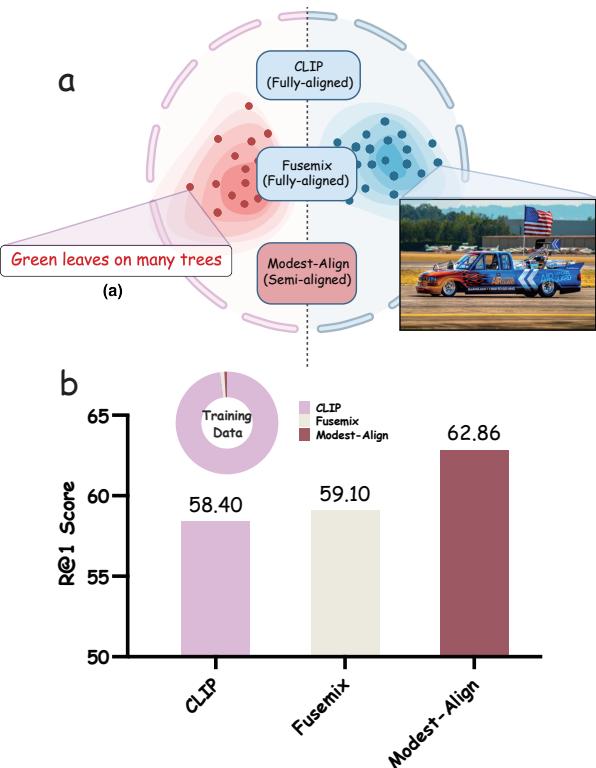

Figure 1: (a) Compared with CLIP and Fusemix, which assume full alignment during training (e.g., "Green leaves on many trees" is treated as a fully matched pair), our Modest-Align handles partially matched or noisy positive pairs by encouraging semi-aligned representations through tailored perturbation and smoothing strategies. (b) On the MS-COCO test dataset (Image → Text retrieval), Modest-Align outperforms both CLIP and Fusemix, despite using only 3.5M training samples—exceeding Fusemix trained on 5M samples (including our 3.5M subset) and CLIP trained on 400M pairs.

unprecedented potential across various application fields, with some applications even attracting mainstream attention (Girdhar et al., 2023; Radford et al., 2021). The cornerstone of multimodal learning is cross-modal alignment, which maps information from multiple modalities, such as text and images, into a unified multimodal vector space (Radford et al., 2021; Alayrac et al., 2022). Researchers have made numerous efforts in cross-modal alignment, with Visual Language Models (VLMs) being particularly representative. VLMs like CLIP (Rad-
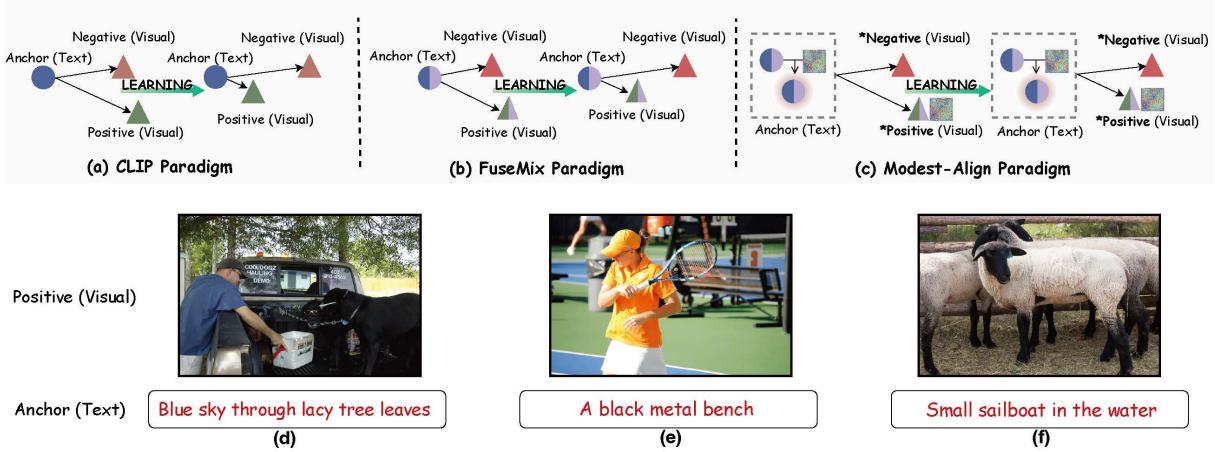
Figure 2: (a-c) Three Paradigms for Cross-modal Alignment: CLIP for contrastive learning, FuseMix for embedding-level mixing, and MODEST-ALIGN for moderating overconfidence and enhancing robustness. Both FuseMix and MODEST-ALIGN derive positive pairs from mixed features, while MODEST-ALIGN further injects random perturbations into visual and textual embeddings to simulate uncertainty. "Positive" and "Negative" denote matched and mismatched pairs, with the gray region indicating uncertainty-aware perturbation. (d–f) Examples of ambiguous samples, including partially matched pairs (d, e) and completely mismatched pairs (f).

ford et al., 2021), which undergo extensive image-text pre-training, excel in image recognition tasks, showcasing the potential of VLMs in establishing effective cross-modal connections.

The success of cross-modal alignment largely relies on large-scale training mechanisms like CLIP, which often require extensive GPU resources and rely on billions of multimodal data pairs (Zhai et al., 2022; Radford et al., 2021; Alayrac et al., 2022). However, the high computational costs are impractical for scenarios with limited computing resources or scarce multimodal data. Therefore, designing a cost-effective and efficient cross-modal alignment framework is crucial. Inspired by Mixup (Zhang et al., 2018), Fusemix introduces an efficient strategy for cross-modal alignment (Vouitsis et al., 2024) by augmenting the latent spaces of pre-trained unimodal encoders, allowing for model creation with significantly reduced data and computational requirements. However, ambiguous samples—whether partially matched or completely unmatched—in datasets with weakly associated (see Figure 1), low-quality image-text pairs can lead to overconfidence and confusion in models, ultimately degrading performance. Moreover, current contrastive learning methods, which rely on single positive examples, exacerbate this issue by further encouraging overconfidence in the presence of ambiguous samples (Vouitsis et al., 2024).

To overcome these issues, we propose MODEST-ALIGN, a cross-modal alignment enhancement method designed to adjust the matching degree of the data and moderate model overconfidence

with a single GPU, incorporating two key components: 1) *Random Perturbation*: This introduces normally distributed perturbations at the visual-text feature level to simulate uncertainty, enhancing the model's generalization capability and helping it learn more robust feature representations. 2) *Embedding Smoothing*: This aims to smooth the model's prediction of output distributions, moderating model overconfidence in positive samples and increasing the smoothness for predictions on uncertain samples, thereby enhancing generalization. By using MODEST-ALIGN to align the latent spaces of pre-trained unimodal encoders, we have developed a highly competitive visual-language (V-L) model. In retrieval tasks, this model not only surpasses existing state-of-the-art (SoTA) methods but also significantly reduces the need for computational resources and data, as detailed in Figure 10. Our study makes two significant contributions:

- Theoretical and empirical analysis reveals that the quality of existing image-text paired datasets is suboptimal, causing VLMs to become confused and overly confident when faced with ambiguous positive pairs (either partially matched (Figure 2 d, e) or completely unmatched (Figure 2 f). This significantly undermines cross-modal alignment.

- We propose a novel V-L alignment method, MODEST-ALIGN, which incorporates *Random Perturbation* to simulate input uncertainty and *Embedding Smoothing* to mitigate overconfidence in positive samples. This

2

MODEST-ALIGN enhances model generalization and robustness, effectively addressing the challenges posed by the suboptimal quality of existing datasets.

2 Related Work

Cross-modal alignment achieves cross-modal synchronization not through direct correspondences between modalities, but implicitly via internal model mechanisms that discern latent semantic connections within the data. The primary objective of these models is to learn a shared latent space capable of jointly encoding multiple modalities, thereby facilitating effective cross-modal alignment (Tan and Bansal, 2019; Li et al., 2020; Yuan et al., 2021; Wang et al., 2022a; Bao et al., 2022; Wang et al., 2022b; Girdhar et al., 2022; Likhosherstov et al., 2023; Zhang et al., 2023; Wu et al., 2023). Image-language alignment is a pivotal area of study in cross-modal alignment, aiming to create universal models capable of interpreting both image and language data. Standard multimodal models usually undergo end-to-end training on image-text pairs. Yet, training these large-scale models from scratch demands substantial computational and data resources, which can restrict scalability. (Arandjelovic and Zisserman, 2017; Lu et al., 2019; Sun et al., 2019; Su et al., 2020; Chen et al., 2020; Li et al., 2021, 2022; Liu et al., 2025).

Pioneered by CLIP and ALIGN (Radford et al., 2021; Jia et al., 2021), this approach uses a dual-encoder architecture, jointly embedding text and images into the same latent space through contrastive target training. 3T aligns text and image encoders with the latent space of a pretrained classifier (Kossen et al., 2023). LiT uses a frozen pretrained image classifier as the image encoder and aligns a text encoder to it (Zhai et al., 2022). Although these methods have seen success, they mostly train one or two encoders from scratch, relying on expensive cross-GPU gradient computations. ImageBind (Girdhar et al., 2023) uses images as anchors to learn a shared latent space across six modalities through contrastive learning, jointly training various modality encoders from scratch. Moreover, the large-scale image-text paired datasets they use, ranging from 400 million to 5 billion pairs, mostly sourced from the internet, are generally not public (Vouitsis et al., 2024). In contrast to these works, Fusemix boosts computational and data efficiency through feature augmentation techniques, using frozen pre-trained unimodal encoders and fewer multimodal paired data, requiring fewer resources (Vouitsis et al., 2024). However, ambiguous positive samples in weakly associated datasets (see Figure 1) lead to model overconfidence and degraded performance, exacerbated by contrastive learning methods that focus on single positive examples (Radford et al., 2021).

Figure 2 compares three alignment paradigms: the CLIP Paradigm, which utilizes contrastive learning to manage data point relationships; the Fusemix Paradigm, which enhances embeddings by mixing features of image-text pairs; and the MODEST-ALIGN Paradigm, which reduces overconfidence through perturbations and embedding smoothing, enhancing generalization and robustness. The MODEST-ALIGN specifically reduces model overconfidence and ensures experiments are computationally and data efficient, requiring only a reasonable amount of GPU resources (Figure 12).

3 Methodology

In this section, we introduce the MODEST-ALIGN framework, designed to facilitate visual-text modal alignment in the latent space while addressing key considerations such as model overconfidence, and computational and data efficiency. MODEST-ALIGN entire process is illustrated in Figure 3.

3.1 Preliminaries

Notation: We define the task of V-L alignment from an alignment perspective. The goal is to learn a shared latent space between visual and textual modal inputs. Formally, given any two data modalities (images $\mathcal{X}$ and text $\mathcal{Y}$), our objective is to learn two networks, $f_X : \mathcal{X} \rightarrow \mathcal{S}$ and $f_Y : \mathcal{Y} \rightarrow \mathcal{S}$, that embed each modality into a shared latent space $\mathcal{S}$.

We take our two encoders as $f_X = F_X \circ A_X$ and $f_Y = G_Y \circ A_Y$. That is, we define $F_X(Frozen)\colon \mathcal{X} \rightarrow \mathcal{U}_{\mathcal{X}}$ and $G_Y(Frozen)\colon \mathcal{Y} \rightarrow \mathcal{U}_{\mathcal{Y}}$, where $\mathcal{U}_{\mathcal{X}}$ and $\mathcal{U}_{\mathcal{Y}}$ are intermediate latent spaces. We then have $A_X(Learnable)\colon \mathcal{U}_{\mathcal{X}} \rightarrow \mathcal{S}$ and $A_Y(Learnable)\colon \mathcal{U}_{\mathcal{Y}} \rightarrow \mathcal{S}$, which we hereafter refer to as V-L adapters. Our insight here is to take both F_X and G_Y as pre-trained unimodal encoders which we keep frozen throughout, and treat our V-L adapters A_X and A_Y as learnable heads for cross-modal alignment. Therefore, we can define our learning objective using the InfoNCE loss function as follows:
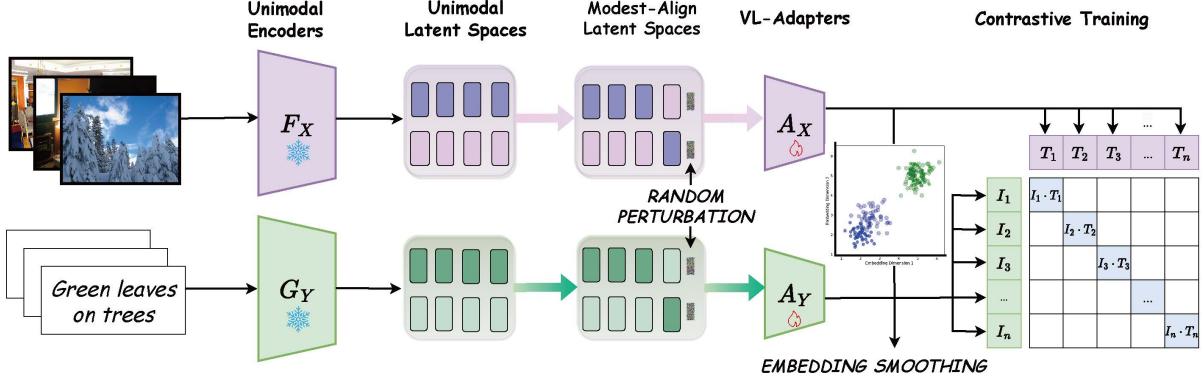
Figure 3: A pipeline of the MODEST-ALIGN showcases the process of aligning the latent spaces of pre-trained unimodal encoders using a fewer dataset of paired data. The unimodal encoders remain frozen, with their latent encodings pre-computed only once for efficiency. In this framework, both Random Perturbation and Embedding Smoothing are applied to each latent space to enhance robustness and reduce model overconfidence. Lightweight V-L adapters are trained to meticulously align these augmented latents into a cohesive, shared latent space, effectively bridging the semantic gap between different modalities.

$$\mathcal{L}_{\text{NCE}} = -\log \frac{\exp(\text{sim}(f_X(\mathcal{X}_i), f_Y(\mathcal{Y}_i))/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(f_X(\mathcal{X}_i), f_Y(\mathcal{Y}_j))/\tau)}, \quad (1)$$

where $\text{sim}(\cdot, \cdot)$ denotes the similarity function, τ is the temperature parameter, and N is the number of all samples. Where $\mathcal{X}_i$ and $\mathcal{Y}_i$ are positive pairs.

Although Eq. (1) provides a standard contrastive formulation to align visual and textual modalities via $f_X = F_X \circ A_X$ and $f_Y = G_Y \circ A_Y$, it assumes sufficient training samples to balance noise and outliers. However, in low-resource scenarios where $N \ll 400M$, many positive pairs $(\mathcal{X}_i, \mathcal{Y}_i)$ may be weakly correlated or noisy, which can lead to overfitting or misalignment during training.

Problem Formulation: To explicitly promote alignment efficiency under limited data, we reformulate the objective from a *unit sample information efficiency* perspective. Specifically, we propose to maximize the alignment quality *per training pair* as follows:

$$\max_{\theta} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{\text{align}} \left(f_X(\mathcal{X}_i), f_Y(\mathcal{Y}_i) \right), \quad (2)$$

where $f_X(\mathcal{X}_i)$ and $f_Y(\mathcal{Y}_i)$ are the projected embeddings in the shared space $\mathcal{S}$ obtained through the visual and textual adapters A_X and A_Y, respectively. $\mathcal{L}_{\text{align}}(\cdot, \cdot)$ measures the quality of alignment between the two modalities—e.g., cosine similarity or contrastive matching score.

Compared to batch-wise InfoNCE, Eq. (2) focuses on per–pair alignment efficiency. Yet, under limited or noisy supervision this objective still inherits an implicit *hard–match* assumption: every designated positive pair $(\mathcal{X}_i, \mathcal{Y}_i)$ is treated as

perfectly aligned, while all other pairs in the mini-batch are considered negatives.

Alignment Distribution. Let

$$s_{ij} = \frac{\langle f_X(\mathcal{X}_i), f_Y(\mathcal{Y}_j) \rangle}{\tau}$$

denote the scaled similarity between the i-th image and the j-th text embedding. The predicted alignment distribution is then:

$$p_{ij} = \frac{\exp(s_{ij})}{\sum_{k=1}^{N} \exp(s_{ik})}, \quad (3)$$

where p_{ij} represents the normalized matching probability in matching $\mathcal{X}_i$ with $\mathcal{Y}_j$.

Failure modes under limited or noisy supervision. With the hard pairwise target $q_{ij} = \mathbf{1}_{\{j=i\}}$, the predicted distribution and its per-sample gradients are:

$$g_{ij} = \frac{\partial \mathcal{L}_i}{\partial s_{ij}} = \begin{cases} p_{ij} - 1, & j = i, \\ p_{ij}, & j \neq i. \end{cases} \quad (4)$$

When the declared positive pair $(\mathcal{X}_i, \mathcal{Y}_i)$ is only weakly aligned, the contrastive objective exhibits three coupled pathologies:

(a) **Over-emphasis on few pairs:** $p_{ii} \ll 1$ still triggers $\max_j p_{ij} \to 1$; the model forces s_{ii} upward and suppresses all s_{ij} ($j \neq i$).

(b) **Entropy collapse:** the prediction entropy

$$H(p_i) = -\sum_{j=1}^{N} p_{ij} \log p_{ij} \longrightarrow 0,$$

yielding an over-confident, brittle alignment distribution.

4

(c) **Unstable gradients / embedding collapse:** the ℓ_1-norm of the gradient vector

$$\left\| \mathbf{g}_i \right\|_1 = |p_{ii} - 1| + \sum_{j \neq i} p_{ij} \approx 1$$

becomes large when p_{ii} is small, resulting in abrupt updates and eventual representation collapse.

3.2 Proposed Solution

The large gradient norm in Eq. (4) is rooted in two factors: (i) the *rigid* supervision q_{ij} that drives $|p_{ii} - q_{ii}| = 1 - p_{ii}$ to its maximum, and (ii) the *sharp* score s_{ii} that contracts the probability mass into a single point. Therefore, an effective remedy must **(a)** soften the supervision signal and **(b)** disperse the score distribution. We next show that both requirements can be satisfied by a *single mathematical modification* of the optimization objective.

Random Perturbation (RP) as Jacobian regularization. We inject an *input–space Gaussian perturbation* to *both* visual and textual embeddings; for the visual branch we have

$$\tilde{f}_X(\mathcal{X}_i) = f_X(\mathcal{X}_i) + \sigma \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, I),$$

and the textual branch is treated in the same manner. The perturbed similarity is defined as $\tilde{s}_{ij} = \langle \tilde{f}_X(\mathcal{X}_i), f_Y(\mathcal{Y}_j) \rangle / \tau$. Taking expectation over ϵ_i we have, by a first–order Taylor expansion,

$$\mathbb{E}_\epsilon[\mathcal{L}_i(\tilde{s}_{i:})] \approx \mathcal{L}_i(s_{i:}) + \frac{\sigma^2}{2} \left\| \nabla_{\tilde{f}_X} \mathcal{L}_i \right\|^2. \quad (5)$$

which is equivalent to adding a *Jacobian norm penalty* to the original loss. Consequently,

$$\left\| \nabla_{\tilde{f}_X} \mathcal{L}_i \right\|_1 \uparrow \text{large} \quad \Longrightarrow \quad \mathcal{L}_{\text{RP}} \text{ increases},$$

forcing the optimiser to seek solutions with smaller gradients. Equation (5) reveals that perturbation training is *equivalent* to adding a Jacobian-norm penalty, directly shrinking the gradient norm in Eq. (4) and hence preventing representation collapse. For inference, the noise is removed to ensure accurate predictions, returning embeddings to their original state.

The large gradient norm in Eq. (4) originates not only from the narrow score landscape but also from the *rigid* pairwise target q_{ij}. This observation motivates a complementary remedy that operates directly on the *target space*:

Embedding Smoothing (ES) as target relaxation. We replace the binary supervision by its convex relaxation:

$$\tilde{q}_{ij} = (1 - \alpha) \mathbf{1}_{\{j=i\}} + \frac{\alpha}{N} \mathbf{1}, \qquad 0 < \alpha < 1, \quad (6)$$

i.e. each sample retains weight $1 - \alpha$ on its annotated match and shares the remaining mass α uniformly with all other pairs. The per-pair loss becomes $\mathcal{L}_i^{\text{ES}} = \text{KL}(\tilde{q}_i \,\|\, p_i)$. ES offers at least three key benefits:

- *Gradient bound.* In Eq. (4) the critical term is now $|p_{ii} - \tilde{q}_{ii}| = \left| p_{ii} - (1 - \alpha + \frac{\alpha}{N}) \right| \leq 1 - \alpha + \frac{\alpha}{N} < 1$, providing an explicit upper limit on the update magnitude.

- *Entropy floor.* Jensen's inequality gives $H(p_i) \geq -\left[(1 - \alpha) \log(1 - \alpha) + \alpha \log(\frac{\alpha}{N})\right]$, guaranteeing a non-degenerate probability distribution and eliminating over-confidence.

- *Convex relaxation.* Because $\tilde{q}_i$ is a convex combination of a delta distribution and the uniform distribution, the KL objective remains convex in p_i, yielding a better behaved optimization landscape.

Confidence-Calibrated Contrastive Loss (CCL). Substituting the perturbed embeddings and smoothed targets into Eq. (2) yields the data-efficient alignment objective (CCL):

$$\mathcal{L}_{\text{ccl}} = \mathbb{E}_\epsilon\left[\frac{1}{N} \sum_{i=1}^{N} \text{KL}(\tilde{q}_i \,\|\, \text{softmax}(\tilde{s}_{i:})) \right] + \lambda \sigma^2, \quad (7)$$

where σ is the perturbation scale and λ is the regularization coefficient. Eq. (7) unifies *Embedding Perturbation* and *Embedding Smoothing* as a principled solution to the instability identified in Eq. (4). The pseudo-code is summarised in Algorithm 1. To our knowledge, this is the first formulation that explicitly integrates uncertainty modeling into a unified objective for cross-modal alignment.

3.3 Pipeline

MODEST-ALIGN adjusts data matching by simulating input uncertainty and moderating model over-confidence in ambiguous positive samples, operating on the latent spaces $\mathcal{U}_\mathcal{X}$ and $\mathcal{U}_\mathcal{Y}$ derived from pre-trained unimodal encoders. 1) The method initially employs unimodal encoders to encode V-L modalities into intermediate latent spaces. 2) Then, it utilizes enhanced features based on Fusemix in

Algorithm 1: Training MODEST-ALIGN

Require: mini-batch $\{(\mathcal{X}_i, \mathcal{Y}_i)\}_{i=1}^{N}$
1: $\mathbf{z}_x \leftarrow F_X(\mathcal{X}); \ \mathbf{z}_y \leftarrow G_Y(\mathcal{Y})$
2: sample $\epsilon, \epsilon' \sim \mathcal{N}(\mathbf{0}, I)$
3: $\tilde{\mathbf{z}}_x \leftarrow \mathbf{z}_x + \sigma\epsilon; \ \tilde{\mathbf{z}}_y \leftarrow \mathbf{z}_y + \sigma\epsilon'$
4: $\tilde{s}_{ij} \leftarrow \langle A_X(\tilde{\mathbf{z}}_{x,i}), A_Y(\tilde{\mathbf{z}}_{y,j}) \rangle / \tau$
5: construct $\tilde{q}_i$ using Eq. (6)
6: $\mathcal{L}$ by Eq. (7)
7: update parameters of A_X and A_Y with AdamW on $\mathcal{L}$

the MODEST-ALIGN latent spaces, incorporating **RP** to adjust data matching by simulating input uncertainty. 3) After training through VL-Adapters, **ES** is applied, aiming to smooth the model's prediction of output distributions, reduce overconfidence in positive pairs, and enhance the smoothness of predictions on uncertain samples. 4) Finally, the smoothed embeddings are used in contrastive learning training, facilitating the learning of two networks, A_X and A_Y, as shown in Figure 3.

MODEST-ALIGN utilizes the existing semantics encoded by unimodal encoders, reducing the reliance on extensive real paired data and simplifying computational requirements. It effectively mitigates the issue of model overconfidence, making the model more "modest" and robust, thereby optimizing cross-modal alignment, learning efficiency, and generalization capabilities.

4 Experiments

Baselines and Metrics: We conduct comparisons with several cross-modal alignment methods: Fusemix (Vouitsis et al., 2024), CLIP (Radford et al., 2021), LIT (Zhai et al., 2022), and 3T (Kossen et al., 2024), as shown in Table 1 and Figure 4. It's important to note that large-scale image-text datasets used by models like CLIP, LIT, and 3T, ranging from 400 million to 5 billion pairs, are mostly sourced from the internet and not publicly available, with high computational costs making them impractical for limited-resource scenarios. Therefore, Fusemix, offers a more applicable comparison with MODEST-ALIGN (Figure 4). We evaluate cross-modal alignment with Recall@K: R@1, R@5, and R@10—Recall at the top 1, 5, and 10 retrieved items for both text-to-image and image-to-text tasks (Vouitsis et al., 2024).

Experimental Setup: To minimize computational demands, all our experiments are conducted on a single 24GB NVIDIA 3090 GPU. We pre-compute latents from pre-trained unimodal encoders, which are then discarded, extracting latents for each modality sequentially to avoid loading more than one encoder at a time. For consistency and fair comparison, we use the same unimodal encoders as Fusemix. V-L adapters are parameterized as lightweight MLPs featuring an inverted bottleneck architecture, inspired by previous studies (Lin et al., 2015; Tolstikhin et al., 2021; Bachmann et al., 2023). Each MLP incorporates residual blocks and a default final projection layer with a dimension of 512, embedding each modality into a shared latent space. For the image encoder, we consider DINOv2 (Oquab et al., 2023), and for the text side, we select text encoder with demonstrably semantic latent spaces, specifically BGE (Xiao et al., 2023). Through grid search (see Appendix), we found optimal values of $\sigma = 0.01$ and $\alpha = 0.1$, and our ablation experiments show that a batch size of 10k provides the best performance for MODEST-ALIGN.

Training and Test Datasets: To evaluate the effectiveness of the MODEST-ALIGN for the task of modality alignment, we conducted extensive comparative experiments against SoTA methods across various datasets. following previous works (Chen et al., 2020; Li et al., 2021, 2022, 2023), These training datasets include COCO (Lin et al., 2014a), Visual Genome (VG) (Krishna et al., 2017a), SBU (Ordonez et al., 2011a), and Conceptual Captions 3M (CC3M) (Sharma et al., 2018a). Table 6 provides detailed information about these four datasets. It is noteworthy that the original CC3M dataset, consisting of images stored as internet URLs, currently has only 1.5 million data pairs available. We tested MODEST-ALIGN on the Flickr (Young et al., 2014a) and MS-COCO datasets (Lin et al., 2014a) to benchmark performance in image-text retrieval and assess generalization across scales.

4.1 Results and Analysis

MODEST-ALIGN demonstrates a significant advantage in achieving high performance with substantially lower training costs. Compared to training datasets of similar size, MODEST-ALIGN outperforms both CLIP and Fusemix, achieving an R@1 improvement of 11.42% in text-to-image retrieval and 14.2% in image-to-text retrieval over CLIP (Table 1). Additionally, it surpasses Fusemix with improvements of 5.82% and 7.2% in both tasks, respectively, as detailed in Table 1. Furthermore, despite using significantly smaller training data

Table 1: The performance of SoTA methods and MODEST-ALIGN on different training datasets is assessed on the Flickr30K's 1K test set and MS-COCO's 5K test set, evaluating text-to-image and image-to-text retrieval accuracy using R@1 scores.

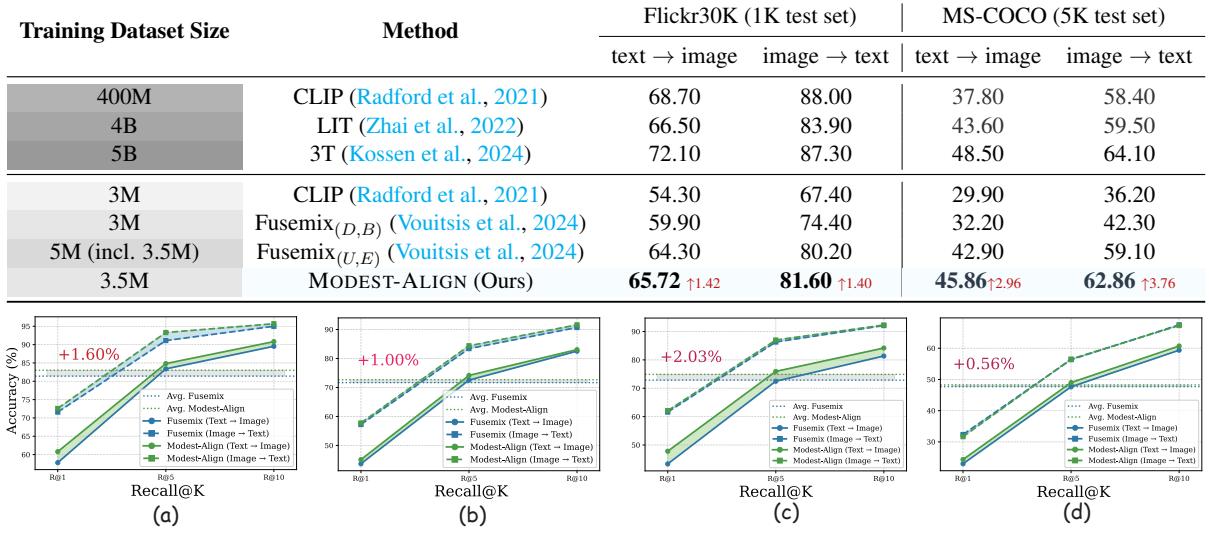| Training Dataset Size	Method	Flickr30K (1K test set)		MS-COCO (5K test set)	
		text → image	image → text	text → image	image → text
400M	CLIP (Radford et al., 2021)	68.70	88.00	37.80	58.40
4B	LIT (Zhai et al., 2022)	66.50	83.90	43.60	59.50
5B	3T (Kossen et al., 2024)	72.10	87.30	48.50	64.10
3M	CLIP (Radford et al., 2021)	54.30	67.40	29.90	36.20
3M	Fusemix$_{(D,B)}$ (Vouitsis et al., 2024)	59.90	74.40	32.20	42.30
5M (incl. 3.5M)	Fusemix$_{(U,E)}$ (Vouitsis et al., 2024)	64.30	80.20	42.90	59.10
3.5M	MODEST-ALIGN (Ours)	**65.72** ↑1.42	**81.60** ↑1.40	**45.86** ↑2.96	**62.86** ↑3.76

Figure 4: Performance of MODEST-ALIGN and FuseMix (Vouitsis et al., 2024) on Flickr30K (a, c) and MS-COCO (b, d) test sets. Evaluated across training sets of varying scale (COCO for (a-b)), SBU for (c-d)), MODEST-ALIGN consistently demonstrates strong generalization and achieves SoTA performance on image retrieval benchmarks. Red indicates average improvement.
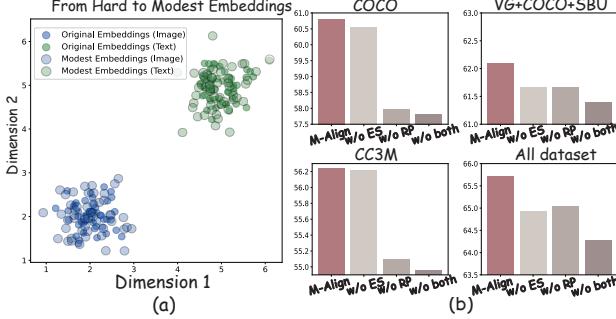

Figure 5: (a) In the original distribution, clusters of image-text pairs are tightly packed, showing high model confidence. After implementing MODEST-ALIGN, the embeddings become more dispersed, indicating reduced confidence in individual labels and a more robust, softer probability distribution that enhances model generalization. (b) The ablation study of two techniques (RP and ES) in Recall@1.

(3.5M pairs), MODEST-ALIGN performs competitively against much larger datasets used by CLIP (400M) and LIT (4B), trailing by only 0.78% in text-to-image and 2.30% in Flickr image-to-text retrieval against LIT. This demonstrates the effectiveness and efficiency of MODEST-ALIGN. We tested MODEST-ALIGN on the MS-COCO, demonstrating its superior performance across various datasets, particularly improving the R@1 score. With 3.5M data, MODEST-ALIGN outperforms CLIP on MS-COCO, requiring 100 times less data and over 600 times less training time, compared to CLIP's 3000 GPU days and 400M data (Figure 12). Classification results of MODEST-ALIGN versus CLIP are reported in the Appendix.

As shown in Figure 4, MODEST-ALIGN consis-

tently outperforms Fusemix across various dataset sizes and settings in image-text retrieval tasks. On the COCO ($560K$ pairs), MODEST-ALIGN achieves a significant improvement in text-to-image retrieval, with a 3% higher R@1 score compared to Fusemix. On SBU ($840K$ pairs), MODEST-ALIGN surpasses Fusemix by over 4.48% in R@1 for text-to-image tasks. Furthermore, on a combined training configuration of four datasets (VG+COCO+SBU+CC3M, totaling $3.5M$ pairs), MODEST-ALIGN achieves a significant improvement of over 1.44% in R@1 for text-to-image retrieval. Besides, we tested on the MS-COCO, demonstrating its superior performance across various datasets. These results highlight MODEST-ALIGN's robust generalization capabilities and superior performance over the SoTA method.

Using CLIP-ViT/B-32 (Radford et al., 2021), we assessed image-text quality and found that most pairs are only weakly aligned: 93.5 % of CC3M, 91.9 % of COCO, 94.4 % of SBU, and 99.1 % of VG have cosine scores < 35 (Table 6, Figure 11). COCO's comparatively higher alignment explains its stronger downstream results (Figure 11). Such low-quality, ambiguous pairs drive over-confidence in standard contrastive training and mislead inference, motivating MODEST-ALIGN.

4.2 Ablation Study

Effect of RP: To validate the effectiveness of RP, we conducted ablation experiments for RP. As
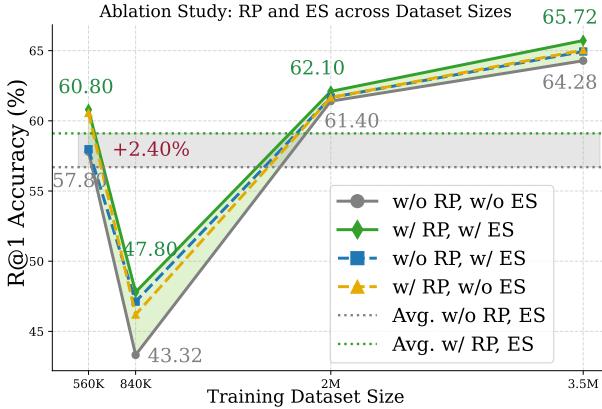
7

Figure 6: Ablation experiments of different techniques, including RP and ES, are conducted across various datasets, measuring text-to-image retrieval performance (R@1).

Table 2: Ablation study on ES variants across datasets (R@1). ✗, denotes training without ES, Dynamic uses adaptive α, and ✓, indicates default α.

| Dataset Size | Different Type of ES | | | | | |
| | text → image | | | image → text | | |
	✗	Dynamic	✓	✗	Dynamic	✓
560K	57.80	60.22	**60.56**	71.60	72.70	**73.50**
820K	51.66	52.50	**53.80**	66.90	68.10	**69.20**
840K	43.32	45.46	**47.12**	61.60	63.60	**62.80**
2M	61.40	61.32	**61.66**	77.20	77.90	**78.40**
3.5M	64.28	65.14	64.94	81.20	81.50	**82.50**

Table 3: Impact of dataset quality and size on Flickr30K retrieval. We compare Fusemix (light red) and MODEST-ALIGN (light blue) under varying data conditions. Smaller, high-quality datasets outperform larger but noisier ones, highlighting the importance of data quality.

| Size | Dataset | text → image | | | image → text | | |
		R@1	R@5	R@10	R@1	R@5	R@10
840K	SBU	43.32	72.48	81.36	61.60	86.30	92.10
560K	COCO	57.80	83.38	89.54	71.60	91.10	95.00
840K	SBU	47.80	75.94	84.18	62.10	87.00	92.30
560K	COCO	60.80	84.82	90.82	72.60	93.30	95.70

shown in Figure 6, adding RP consistently improves performance across all datasets. On the COCO, adding RP in two settings—with and without ES—results in an R@1 score increase from 57.98% to 60.8%, a 2.82% improvement, and from 57.8% to 60.56%, a 2.76% improvement. Similarly, this positive trend in gains from RP is observable across other datasets as well, as shown in Figure 5 (b). These results demonstrate that RP significantly enhances model performance regardless of whether ES techniques are used.

Effect of ES: As shown in Figure 6, ES improves the R@1 on the SBU by 3.8% without RP and by 1.6% with RP. Similar gains from ES are observed across other datasets, regardless of the RP, as shown in Figure 5. ES enhances performance on complex, large-scale datasets by improving generalization across diverse image-text pairs. As shown in Table 2, Default ES outperforms dynamic α (α is dynamically decayed over training epochs from 0.1 to 0.01 to progressively reduce supervision rigidity), achieving a 0.34% improvement in R@1 for text-to-image retrieval and 0.8% for image-to-text on COCO. These gains are consistent across datasets, with default ES proving more effective in optimizing retrieval tasks and handling complex data. Both smoothing methods outperform models without ES, highlighting its effectiveness.

Impact of Dataset Quality: Human-annotated COCO ($560K$) outperforms larger SBU (($840K$) in retrieval, achieving 57.80% vs. 43.32% R@1, regardless of the method used (Table 3). This shows that dataset curation is more impactful than simply increasing size. MODEST-ALIGN's ability to adjust image-text matching is valuable, especially since large-scale curation is impractical. As shown in

Figure 5 (a), MODEST-ALIGN softens hard embeddings into smoother, more dispersed ones, lowering overconfidence and improving generalization.

4.3 Discussion

ES, inspired by label smoothing, adapts its core idea to the embedding space for cross-modal alignment. By smoothing the similarity distribution between positive and negative pairs, ES mitigates overconfidence and improves robustness to weakly correlated samples. RP's uniqueness lies in its primary goal to simulate uncertainty in data pairs for cross-modal alignment tasks, enhancing model robustness. RP is designed for the embedding space, helping the model learn more "modestly" on weakly correlated datasets. MODEST-ALIGN is the first to unify them under a cohesive objective tailored for cross-modal uncertainty, enabling robust contrastive learning over noisy or weakly aligned data. This combination yields a principled loss tailored for robust alignment under uncertainty.

5 Conclusion

In this work, we propose MODEST-ALIGN, a data- and compute-efficient cross-modal alignment method that mitigates overconfidence, strengthens latent pairwise associations, and leverages pretrained unimodal encoders for guidance. Validated across multiple datasets, MODEST-ALIGN has consistently demonstrated its robust capability to align V-L models better.

Limitations

Trained on a single GPU, MODEST-ALIGN consistently improved alignment across all datasets, with larger gains on lower-quality data. One limitation of MODEST-ALIGN is that its performance on extremely large-scale datasets remains unclear, as training such datasets is impractical under limited GPU resources. Additionally, it is uncertain how much MODEST-ALIGN would benefit datasets with very high-quality, perfectly matched image-text pairs, because such data is rare.

For future developments, MODEST-ALIGN could incorporate data quality assessments to dynamically adjust model confidence, applying stricter constraints on lower quality data and more lenient ones on higher quality data to effectively manage model overconfidence.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736.

Relja Arandjelovic and Andrew Zisserman. 2017. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*.

Gregor Bachmann, Sotiris Anagnostidis, and Thomas Hofmann. 2023. Scaling MLPs: A Tale of Inductive Bias. *arXiv:2306.13575*.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. 2022. VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts. In *Advances in Neural Information Processing Systems*, volume 35, pages 32897–32912.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning. In *Computer Vision – ECCV 2020*, pages 104–120.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. ImageBind: One Embedding Space To Bind Them All. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190.

Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. 2022. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16102–16112.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 4904–4916. PMLR.

Jannik Kossen, Mark Collier, Basil Mustafa, Xiao Wang, Xiaohua Zhai, Lucas Beyer, Andreas Steiner, Jesse Berent, Rodolphe Jenatton, and Effrosyni Kokiopoulou. 2024. Three towers: Flexible contrastive learning with pretrained image models. *Advances in Neural Information Processing Systems*, 36.

Jannik Kossen, Mark Collier, Basil Mustafa, Xiao Wang, Xiaohua Zhai, Lucas Beyer, Andreas Steiner, Jesse Berent, Rodolphe Jenatton, and Efi Kokiopoulou. 2023. Three towers: Flexible contrastive learning with pretrained image models. *arXiv:2305.16999*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael S. Bernstein, and Fei-Fei Li. 2017a. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017b. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pretraining for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 12888–12900. PMLR.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems*, volume 34, pages 9694–9705.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *Computer Vision – ECCV 2020*, pages 121–137.

Valerii Likhosherstov, Anurag Arnab, Krzysztof Marcin Choromanski, Mario Lucic, Yi Tay, and Mostafa Dehghani. 2023. PolyViT: Co-training Vision Transformers on Images, Videos and Audio. *Transactions on Machine Learning Research*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014a. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, pages 740–755. Springer International Publishing.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014b. 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.

Zhouhan Lin, Roland Memisevic, and Kishore Konda. 2015. How far can we go without convolution: Improving fully-connected networks. *arXiv:1511.02580*.

Jiaxiang Liu, Jin Hao, Hangzheng Lin, Wei Pan, Jianfei Yang, Yang Feng, Gaoang Wang, Jin Li, Zuolin Jin, Zhihe Zhao, et al. 2023a. Deep learning-enabled 3d multimodal fusion of cone-beam ct and intraoral mesh scans for clinically applicable tooth-bone reconstruction. *Patterns*, 4(9).

Jiaxiang Liu, Tianxiang Hu, Jiawei Du, Ruiyuan Zhang, Joey Tianyi Zhou, and Zuozhu Liu. 2025. Kpl: Training-free medical knowledge mining of vision-language models. *arXiv preprint arXiv:2501.11231*.

Jiaxiang Liu, Tianxiang Hu, Huimin Xiong, Jiawei Du, Yang Feng, Jian Wu, Joey Zhou, and Zuozhu Liu. 2024a. Vpl: Visual proxy learning framework for zero-shot medical image diagnosis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9978–9992.

Jiaxiang Liu, Tianxiang Hu, Yan Zhang, Yang Feng, Jin Hao, Junhui Lv, and Zuozhu Liu. 2023b. Parameter-efficient transfer learning for medical visual question answering. *IEEE Transactions on Emerging Topics in Computational Intelligence*.

Jiaxiang Liu, Tianxiang Hu, Yan Zhang, Xiaotang Gai, Yang Feng, and Zuozhu Liu. 2023c. A chatgpt aided explainable framework for zero-shot medical image diagnosis. *arXiv preprint arXiv:2307.01981*.

Jiaxiang Liu, Yuan Wang, Jiawei Du, Joey Zhou, and Zuozhu Liu. 2024b. Medcot: Medical chain of thought via hierarchical expert. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17371–17389.

Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Ilya Loshchilov and Frank Hutter. 2018. Fixing weight decay regularization in adam.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems*, volume 32.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Russel Galuba, Wojciech Howes, Po-Yao Huang, Li Shang-Wen, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2023. DINOv2: Learning robust visual features without supervision. *arXiv:2304.07193*.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011a. Im2Text: Describing Images Using 1 Million Captioned Photographs. In *Advances in Neural Information Processing Systems*, volume 24.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011b. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24.

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. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 8748–8763. PMLR.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018a. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018b. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pretraining of Generic Visual-Linguistic Representations. In *International Conference on Learning Representations*.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. VideoBERT: A Joint Model for Video and Language Representation Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5100–5111.

Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. 2021. MLP-Mixer: An all-MLP architecture for vision. In *Advances in Neural Information Processing Systems*, volume 34, pages 24261–24272.

Noël Vouitsis, Zhaoyan Liu, Satya Krishna Gorti, Valentin Villecroze, Jesse C Cresswell, Guangwei Yu, Gabriel Loaiza-Ganem, and Maksims Volkovs. 2024. Data-efficient multimodal fusion on a single gpu. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27239–27251.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. 2022a. InternVideo: General video foundation models via generative and discriminative learning. *arXiv:2212.03191*.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2022b. SimVLM: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations*.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023. NExT-GPT: Any-to-any multimodal LLM. *arXiv:2309.05519*.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. 2023. C-pack: Packaged resources to advance general chinese embedding. *arXiv:2309.07597*.

Yong Yang, Jiaxiang Liu, Shuying Huang, Weiguo Wan, Wenying Wen, and Juwei Guan. 2021. Infrared and visible image fusion via texture conditional generative adversarial network. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(12):4771–4783.

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

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

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. 2021. Florence: A new foundation model for computer vision. *arXiv:2111.11432*.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. LiT: Zero-Shot Transfer With Locked-Image Text Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*.

Yiyuan Zhang, Kaixiong Gong, Kaipeng Zhang, Hongsheng Li, Yu Qiao, Wanli Ouyang, and Xiangyu Yue. 2023. Meta-transformer: A unified framework for multimodal learning. *arXiv:2307.10802*.

A Appendix

In this section, we present additional implementation details, experiment results, theoretical analysis, pseudo code and supplements. The content structure is outlined as follows:

Figure 7: Ablation for σ on R@1. σ Regulates random perturbation intensity, introducing noise to simulate uncertainty in embeddings and reduce overconfidence from weakly correlated data. Smaller σ ensures stability, while larger σ risks excessive uncertainty.

Figure 8: Ablation for α on R@1. α Controls embedding smoothing to reduce overconfidence when handling weakly correlated positive samples. It smooths the embedding space, making similarity predictions for image-text pairs more robust. An appropriate α enhances generalization while minimizing reliance on perfectly matched positive samples.

Table 4: Ablation for N on Different Datasets. To systematically investigate the impact of batch size on performance, we conduct ablation experiments with varying batch sizes ($1k, 2k, 5k, 10k, 15k$). We evaluate how these changes affect MODEST-ALIGN's performance, especially on text-to-image and image retrieval metrics, to better understand the role of batch size in embedding smoothing. Our results indicate that a batch size of 10k offers optimal performance.

Ablation for N	$1k$	$2k$	$5k$	$10k$	$15k$
T→I (R@1)	58.30	59.08	59.30	**60.80**	59.38
I→T (R@1)	70.00	71.70	**73.10**	72.60	73.00
Average	64.15	65.39	66.20	**66.70**	66.19

A.1 Supplementary Experimental Results

A.1.1 Results of MODEST-ALIGN in Zero-shot Classification

MODEST-ALIGN aims to optimize cross-modal alignment while minimizing data and resource demands. Evaluating zero-shot classification is essential to further validate its generality.

To evaluate MODEST-ALIGN's out-of-distribution (OoD) performance on tasks like zero-shot classification and understand its cross-task applicability, we analyzed its performance
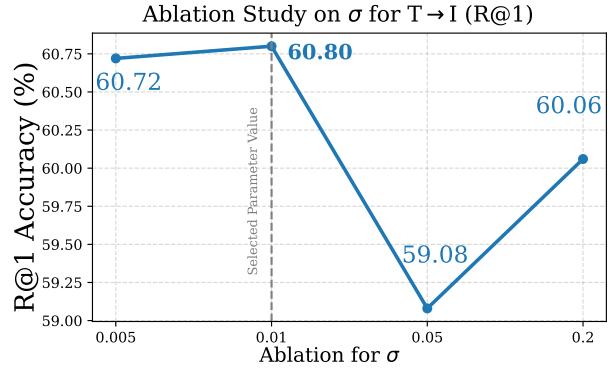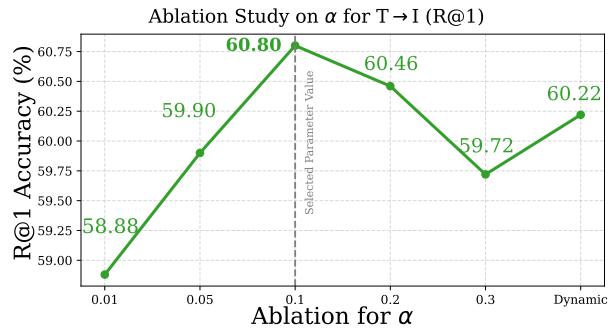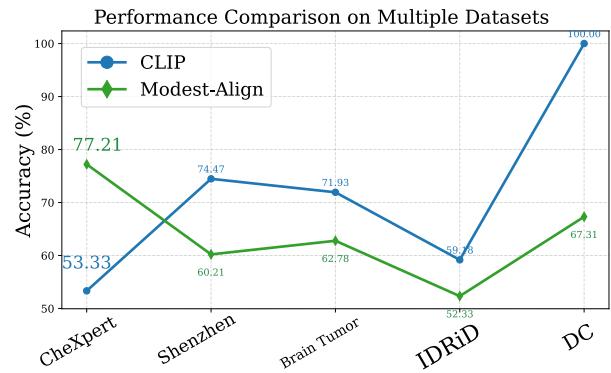

Figure 9: Comparison between CLIP and MODEST-ALIGN on Various Datasets in Zero-shot Classification
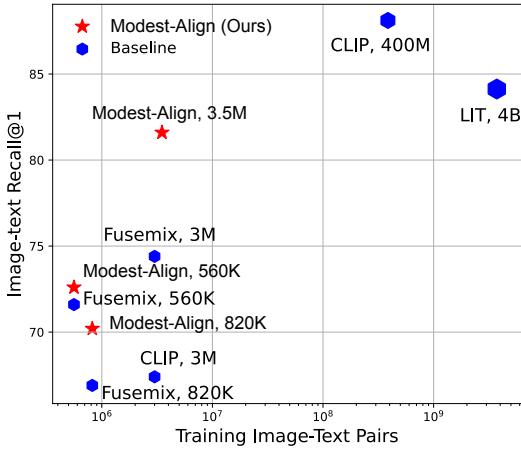
Figure 10: Image-to-Text retrieval performance on the Flickr30K test set (Young et al., 2014b) is plotted against the number of training pairs on a log-scale x-axis, illustrating how training volume impacts effectiveness.

on diverse medical datasets (CheXpert, Shenzhen, IDRiD, Brain Tumor (Liu et al., 2023c)). Specifically, 50 images from different categories were selected as subsets. As shown in the table, MODEST-ALIGN achieved a 77.21% accuracy on the CheXpert, significantly outperforming CLIP's 53.33% accuracy. On other datasets, MODEST-ALIGN's results were close to CLIP's classification performance (Figure 9).

To validate MODEST-ALIGN's accuracy on natural images, we tested it on Kaggle's open-source D&C dataset. Results showed that CLIP achieved 100.00% accuracy, significantly outperforming MODEST-ALIGN's 67.31%. These results highlight CLIP's strength in its advantage of pretraining on a 400M dataset. However, MODEST-ALIGN demonstrated effectiveness across multiple datasets in zero-shot classification tasks, underscoring its generalization capabilities.

With only 3.5M data, MODEST-ALIGN achieves performance comparable to or even surpassing CLIP on certain datasets, while using 100 times less data and over 600 times less training time than CLIP, which requires 3000 GPU days and 400M training data. This highlights the importance of MODEST-ALIGN's applicability in resource-constrained environments.

A.1.2 Parameter Search for MODEST-ALIGN

To systematically investigate the impact of batch size on performance, we conduct ablation experiments with varying batch sizes (1k, 2k, 5k, 10k,

and 15k). We evaluate how these changes affect MODEST-ALIGN's performance, especially on text-to-image and image retrieval metrics, to better understand the role of batch size in embedding smoothing. Our results indicate that a batch size of 10k offers optimal performance (Table 4).

Impact of parameter σ: We fix $\lambda = 1$ and treat σ as the sole hyper-parameter in Equation 7. As the control parameter for random perturbation intensity, σ influences the model's robustness in the embedding space. Experimental results indicate: Smaller σ values (e.g., 0.005) maintain embedding stability but may be insufficient for handling weakly correlated samples. Larger σ values (e.g., 0.05 or higher) enhance the model's adaptability to noise but may overly disturb the feature space, negatively impacting alignment accuracy (Figure 7).

Impact of parameter α: As the control parameter for embedding smoothing, α influences the distribution smoothness between positive and negative pairs. An appropriate α value effectively mitigates overconfidence issues and enhances generalization: Lower α values (e.g., 0.05) may fail to adequately reduce the model's overconfidence. Higher α values (e.g., 0.3) may make the model less confident even for fully matched positive samples, causing interference with positive pair alignment (Figure 8).

A.2 Assessing the Match Quality of Image-text Datasets with CLIP

A.2.1 Similarity Calculation

In the context of contrastive learning models such as CLIP, the similarity produced for a given image-text pair is closely related to the cosine similarity of their respective embeddings, modulated by a temperature scaling factor τ. Specifically, let $Similarity(i, j)$ denote the similarity score for image i and text j. This score can be mathematically expressed as:

$$Similarity(i,j) = \frac{\text{image\_embedding}(i) \cdot \text{text\_embedding}(j)}{\|\text{image\_embedding}(i)\|\|\text{text\_embedding}(j)\|} \times \frac{1}{\tau}, \tag{1}$$

where τ is the temperature coefficient, which is set to 0.01. Consequently, when multiplied by the temperature coefficient $\frac{1}{\tau}$, the similarity score will be constrained within the new range.

A.2.2 Similarity in Four Datasets

To assess the match quality of image-text datasets in the main text, we conducted similarity calculation experiments on four widely used image-text datasets: COCO, CC3M, SBU, and VG. These

Green leaves on many trees | Blue sky through lacy tree leaves | A black metal bench | Small sailboat in the water

(a)   (b)   (c)   (d)

CC3M   COCO   SBU   VG
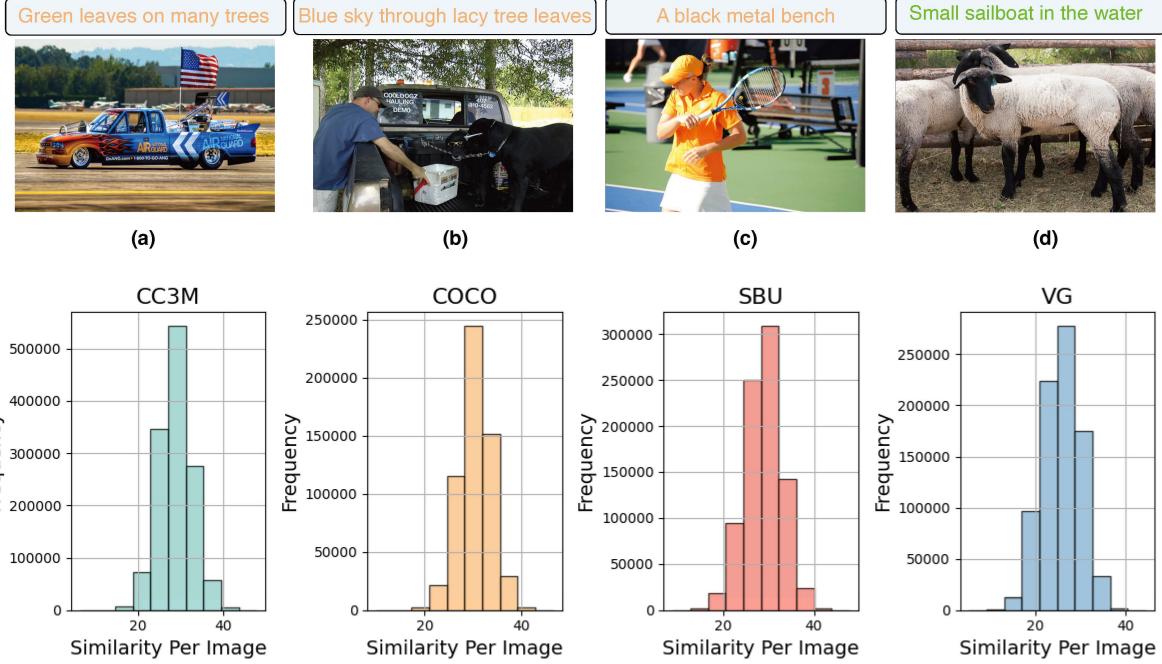
Frequency — Similarity Per Image

Figure 11: The upper figure mentions "ambiguous samples" in datasets, including partially matched samples (a, b, c) and completely unmatched samples (d). During contrastive learning training, it can lead to positive pairs not being truly positive. The lower figure use CLIP to compute similarity for four datasets (For details, see Table 6), revealing generally moderate alignment between images and texts (mostly ranging from 20-40%). Notably, the COCO dataset shows higher alignment (indicative of superior data quality), which correlates with relatively higher performance metrics.

Table 5: The performance of MODEST-ALIGN and Fusemix (Vouitsis et al., 2024) on Flickr30K and MS-COCO test sets. By evaluating on multiple datasets with varying sizes and complexities, we show that our MODEST-ALIGN model exhibits strong generalization capabilities and achieves SoTA performance on image retrieval tasks. Bold signifies the best.

| | | Flickr30K (1K test set) | | | | | | MS-COCO (5K test set) | | | | | |
| | | text → image | | | image → text | | | text → image | | | image → text | | |
Training Dataset Size	Method	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
$\approx 560K$	Fusemix	57.80	83.38	89.54	71.60	91.10	95.00	43.68	72.53	82.53	57.22	83.40	90.68
	MODEST-ALIGN (ours)	**60.80**	**84.82**	**90.82**	**72.60**	**93.30**	**95.70**	**45.11**	**74.16**	**83.05**	**57.76**	**84.38**	**91.60**
$\approx 840K$	Fusemix	43.32	72.48	81.36	61.60	86.30	92.10	22.98	47.64	59.37	**32.36**	56.34	**67.50**
	MODEST-ALIGN (ours)	**47.80**	**75.94**	**84.18**	**62.10**	**87.00**	**92.30**	**24.35**	**49.00**	**60.71**	31.66	**56.54**	67.28
$\approx 2M$	Fusemix	61.40	84.82	90.46	77.20	94.40	97.20	43.76	72.74	82.36	60.94	84.72	91.68
	MODEST-ALIGN (ours)	**62.10**	**85.52**	**90.76**	**77.80**	**94.60**	**97.70**	**44.27**	**72.93**	**82.77**	**61.42**	**84.98**	**91.92**
$\approx 3.5M$	Fusemix	64.28	87.60	91.86	81.20	**96.40**	98.20	45.54	73.35	82.73	62.56	**86.38**	**93.04**
	MODEST-ALIGN (ours)	**65.72**	**87.82**	**92.50**	**81.60**	96.00	**98.30**	**45.86**	**73.73**	**83.16**	**62.86**	85.66	92.42

datasets cover a diverse range of image content, from everyday objects to complex scenes, and vary in terms of size and annotation quality. Table 6 presents a summary of these datasets, detailing the number of image-text pairs in each and the ranges of similarity scores calculated for different image-text pairings.

In a similarity analysis, the COCO dataset exhibited the highest mean similarity per image at 30.48, indicating a strong and consistent association between images and text. The CC3M and SBU datasets demonstrated similar mean similarities of 28.80, reflecting comparable levels of alignment quality. This highlights how dataset structure and complexity significantly influence model perfor-

mance in multimodal tasks, as shown in Figure 11 (b). Table 6 shows the distribution of similarity for Image-Text pairs across four datasets. It can be observed that the values are concentrated around the 30 range, indicating that when CLIP is used as a scoring model, the resulting similarity tends to cluster within a relatively modest range.

A.3 Theoretical Analysis for Embedding Smoothing

Embedding Smoothing effectively increases the entropy of the target distributions by assigning non-zero probabilities to all classes (examples in the batch). This reduction in confidence prevents the model from becoming overly reliant on specific

Table 6: Similarity Per Image-Text Across Ranges for Different Datasets. The table shows the count of similarity scores within specific similarity ranges for each dataset (CC3M, COCO, SBU, and VG). Each row corresponds to a dataset, and each column represents a range of similarity values.

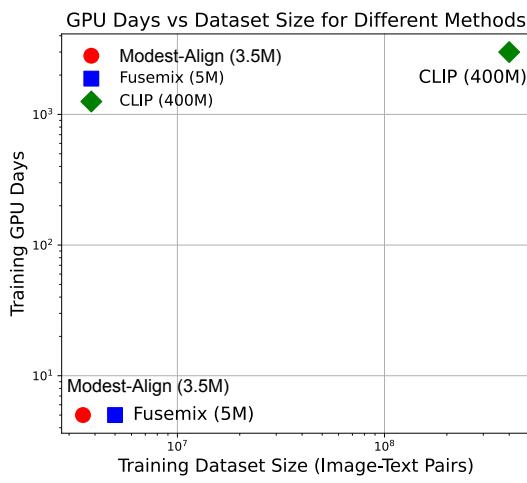|	[5,10)	[10, 15)	[15, 20)	[20, 25)	[25, 30)	[30, 35)	[35, 40)	[40, 45)	[45, 50)
CC3M	11	400	14877	189355	616912	408777	70836	3389	32
COCO	1	29	1187	27566	216780	275238	44607	1331	8
SBU	15	647	15268	127793	365869	284130	45279	1702	24
VG	39	3149	68441	267798	342860	132259	7151	77	0

Figure 12: In terms of training efficiency, CLIP requires 3000 GPU days for training on 400 million data pairs, while MODEST-ALIGN needs only approximately 5 GPU days for 3.5 million data pairs. Our proposed MODEST-ALIGN aims to enhance computational and data efficiency for modal alignment in resource-efficient settings. It has outperformed state-of-the-art methods on public datasets, notably improving the R@1 score in retrieval tasks. With only 3.5M data, MODEST-ALIGN exceeds CLIP's performance on the MS COCO dataset and does so with 100 times less data and over 600 times less training time than CLIP, which requires 3000 GPU days and 400M training data.

training examples, thereby enhancing its robustness. The inclusion of the smoothing parameter α allows for control over the degree of smoothing applied, enabling a balance between model confidence and generalization ability. To provide a theoretical understanding of how Embedding Smoothing improves generalization, we analyze its impact on the loss function and the model's predictions.

In standard contrastive learning without smoothing, the loss for a positive pair is:

$$\mathcal{L}_{\text{pos}} = -\log \frac{\exp\left(\text{sim}\left(A_X(z_x), A_Y(z_y)\right)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\text{sim}\left(A_X(z_x), A_Y(z_j)\right)/\tau\right)}, \quad (2)$$

This loss encourages the model to maximize the similarity between positive pairs and minimize it between negative pairs. However, it can lead to

overconfident predictions, as the model focuses heavily on the positive pair. With Embedding Smoothing, the loss incorporates the smoothed target distribution $\tilde{y}$, and the KL divergence becomes:

$$\mathcal{L}\left(A_X(z_x), \tilde{Y}; Z_Y\right) = -\sum_{i=1}^{N} \tilde{y}_i \log p_i, \quad (3)$$

where p_i is the predicted probability for the i-th example in the batch:

$$p_i = \frac{\exp\left(\text{sim}\left(A_X(z_x), A_Y(z_i)\right)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\text{sim}\left(A_X(z_x), A_Y(z_j)\right)/\tau\right)}, \quad (4)$$

By assigning non-zero probabilities $\tilde{y}_i$ to all classes, the loss function penalizes the model not only for the positive pair but also for negative pairs, albeit to a lesser extent. This encourages the model to produce a probability distribution that is more uniform and less confident.

Analyzing from the perspective of information entropy: The entropy $H(\tilde{y})$ of the smoothed target distribution is higher than that of a one-hot distribution. The entropy of $\tilde{y}$ is:

$$H(\tilde{y}) = -\left((1-\alpha)\log(1-\alpha) + (N-1)\left(\frac{\alpha}{N-1}\log\frac{\alpha}{N-1}\right)\right),$$

(5)

Higher entropy in the target distribution leads to smoother gradients during training, which can prevent the model from fitting noise in the training data. This smoothing effect acts as a form of regularization, reducing overfitting.

Reduction in Overconfident Predictions: Embedding Smoothing reduces the Kullback-Leibler divergence between the predicted distribution p and the uniform distribution u, where $u_i = \frac{1}{N}$:

$$\text{KL}(u\|p) = \sum_{i=1}^{N} u_i \log\frac{u_i}{p_i}, \quad (6)$$

By making p closer to $\tilde{y}$, which has higher entropy, the model's predictions become less confident. This can be beneficial because overconfident

predictions on training data often lead to poor generalization on unseen data.

Connection to Label Smoothing Theory: Embedding Smoothing in our context is analogous to label smoothing in classification tasks. Previous works have shown that label smoothing has the following effects: 1). Margin Maximization: It implicitly increases the decision margin between classes, which can improve generalization. 2). Penalization of Confident Wrong Predictions: By smoothing the targets, the loss function penalizes overconfident incorrect predictions more heavily.

Besides, consider the gradient of the loss with respect to the logits z:

$$\frac{\partial \mathcal{L}}{\partial z_i} = p_i - \tilde{y}_i, \qquad (7)$$

When using Embedding Smoothing, $\tilde{y}_i$ is never exactly 0 or 1. This means that the gradients are non-zero for all classes, encouraging the model to adjust its predictions across all examples in the batch. This leads to more generalized feature representations. By incorporating Embedding Smoothing into the loss function, we introduce a regularization effect that enhances the model's generalization capabilities. The smoothing parameter α provides a mechanism to control this effect, allowing for a trade-off between fitting the training data and maintaining robustness to unseen data. This theoretical understanding aligns with our experimental observations, where models trained with Embedding Smoothing demonstrate improved performance on validation datasets.

A.4 Theoretical Analysis for Random Perturbation

Why Choose Gaussian noise?: This paper introduces Gaussian noise as a perturbation in MODEST-ALIGN for several reasons: 1). Well-Defined Mathematical Properties: Gaussian distribution exhibits continuous and smooth probability density functions across the real number line, facilitating theoretical analysis and calculations. 2). Zero-Mean Symmetry: By choosing a Gaussian distribution with a mean of zero, the added noise is symmetrically balanced around zero, introducing no systematic bias and only increasing the variance, thereby preserving the expected value of embeddings. 3). Adjustable Perturbation Intensity: The standard deviation of the Gaussian distribution can be precisely controlled, allowing for careful calibration of noise

intensity. This flexibility is crucial for introducing an appropriate level of uncertainty to enhance model robustness. 4). Alignment with Natural Phenomena: According to the Central Limit Theorem, the sum of many independent random variables tends toward a Gaussian distribution. Thus, Gaussian noise effectively simulates random disturbances or measurement errors prevalent in natural and engineering contexts. 5). Facilitation of Optimization and Training: In deep learning, incorporating Gaussian noise helps to smooth the loss function landscape, avoiding local minima and promoting more effective training processes.

In summary, choosing Gaussian noise as the source of perturbation provides theoretical soundness and practical convenience, aiding in the development of more robust feature representations, preventing overfitting, and enhancing the generalization capabilities of models.

Loss Function with Perturbed Embeddings: The perturbed embeddings are used in the training loss, specifically in contrastive learning with the InfoNCE loss. The objective is to maximize the similarity between the noisy visual and textual embeddings:

$$\mathcal{L}_{\text{NCE}} = -\log \frac{\exp(\text{sim}(\tilde{\mathbf{z}}_v, \tilde{\mathbf{z}}_t)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(\tilde{\mathbf{z}}_v, \tilde{\mathbf{z}}_t^j)/\tau)}, \quad (8)$$

where $\text{sim}(\cdot, \cdot)$ represents a similarity function, τ is a temperature parameter, and N is the number of negative samples. This formulation ensures that the model learns representations that are invariant to noise, thus improving generalization.

Noise and Regularization Effect: To understand the impact of Gaussian noise on regularization, we first look at the expected value of the perturbed embeddings. Since the noise is zero-mean, the expected value of the perturbed embeddings is identical to the original embeddings:

$$\mathbb{E}[\tilde{\mathbf{z}}_v] = \mathbb{E}[\mathbf{z}_v + \sigma \epsilon_v] = \mathbf{z}_v, \qquad (9)$$

However, the variance of the perturbed embeddings increases due to the added noise. The variance of the perturbed embeddings can be calculated as:

$$\text{Var}[\tilde{\mathbf{z}}_v] = \text{Var}[\mathbf{z}_v + \sigma \epsilon_v] = \text{Var}[\mathbf{z}_v] + \sigma^2 \text{Var}[\epsilon_v], \qquad (10)$$

Given that $\text{Var}[\epsilon_v] = I$, where I is the identity matrix, the total variance of the perturbed embeddings becomes:

$$\text{Var}_{\text{total}} = \text{Var}[\mathbf{z}_v] + \sigma^2 I, \qquad (11)$$

The additional term $\sigma^2 I$ acts as a regularizer, which spreads out the embeddings and prevents the model from becoming overconfident in its predictions.

Minimizing the Generalization Error: The added noise effectively smooths the decision boundary of the model, which reduces overfitting. By introducing noise, we minimize the generalization error. Assuming the model's prediction function is $f(\mathbf{z})$ and the true function is $f^*(\mathbf{z})$, the goal is to minimize the expected generalization error:

$$\mathbb{E}_{\mathbf{z}}[(f(\mathbf{z}) - f^*(\mathbf{z}))^2],$$

With noise perturbation, the variance in the embeddings increases, which forces the model to learn smoother decision boundaries. The regularization effect introduced by the noise helps bind the generalization error:

$$\mathbb{E}_{\mathbf{z}}[(f(\mathbf{z}) - f^*(\mathbf{z}))^2] \leq \text{Var}_{\text{total}} = \text{Var}[\mathbf{z}_v] + \sigma^2, \qquad (12)$$

Thus, the noise helps control the generalization error by ensuring that the model does not overfit to specific features of the training data, which is especially important in cases where the training data contains noise or is limited in size.

Noise-Induced Gradient Regularization: We can also analyze the effect of noise on the gradient of the loss function. Given a loss function $\mathcal{L}(\mathbf{z}_v, \mathbf{z}_t)$, the gradient with respect to the perturbed embeddings can be expressed as:

$$\nabla_{\tilde{\mathbf{z}}_v}\mathcal{L}(\tilde{\mathbf{z}}_v, \tilde{\mathbf{z}}_t) = \nabla_{\mathbf{z}_v}\mathcal{L}(\mathbf{z}_v, \mathbf{z}_t) + \sigma \cdot \nabla_{\epsilon_v}\mathcal{L}(\tilde{\mathbf{z}}_v, \tilde{\mathbf{z}}_t), \qquad (13)$$

The second term, $\sigma \cdot \nabla_{\epsilon_v}\mathcal{L}(\tilde{\mathbf{z}}_v, \tilde{\mathbf{z}}_t)$, acts as a regularizer that prevents the gradient from becoming too large. The gradient is smoothed by the presence of noise, which further prevents overfitting and encourages the model to learn more generalizable patterns.

Inference Phase: During the inference phase, we remove the Gaussian noise to ensure accurate predictions on unseen data. The embeddings revert to their original clean form:

$$\mathbf{z}_v = F(\mathbf{x}_v), \quad \mathbf{z}_t = G(\mathbf{x}_t), \qquad (14)$$

Without the added noise, the model makes precise predictions based on the robust features it learned during training. Therefore, by adding Gaussian noise to the embeddings during training, we introduce a form of regularization that improves the generalization ability of the model. The noise prevents overfitting by increasing the variance of the embeddings, ensuring that the model learns smoother decision boundaries. This leads to better performance on unseen data and helps minimize the generalization error. The noise-induced gradient regularization further contributes to preventing the model from overfitting to the training data, making it more robust in real-world applications.

A.5 Implementation Details

For all experiments, we use the AdamW (Loshchilov and Hutter, 2018) optimizer during training. We perform learning rate warmup by linearly increasing the learning rate from 10^{-6} to 10^{-3}. We then decay the learning rate using a cosine schedule (Loshchilov and Hutter, 2016). We use a depth of 4 for both V-L adapters which we train for 500 epochs with a batch size of $Batch = $ 10K. We set the learning rate as lr$= 10^{-3}$ and use weight decay of 0.1 during optimization. Additionally, τ_t is set to 0.07. The image encoder is DINOv2 ViT-G/14, and for the text side, the text encoder is the BGE large version. To evaluate the effectiveness of the MODEST-ALIGN method for the task of modality alignment, we conducted extensive comparative experiments against SoTA methods across various datasets. These datasets include COCO (Lin et al., 2014b), VG (Krishna et al., 2017b), SBU (Ordonez et al., 2011b), and CC3M (Sharma et al., 2018b). Table 6 provides detailed information about these four datasets. Additionally, we compared different schemes such as Fusemix, CLIP, and LIT. Utilizing a single NVIDIA 3090 GPU for training, MODEST-ALIGN demonstrated SoTA performance across datasets of varying sizes.