FANoise: Singular Value-Adaptive Noise Modulation for Robust Multimodal Representation Learning

Jiaoyang Li*, Jun Fang*, Tianhao Gao, Xiaohui Zhang, Zhiyuan Liu, Chao Liu†, Pengzhang Liu, Qixia Jiang

JD, Retail, Beijing, China

{lijiaoyang7, fangjun8, gaotianhao1, zhangxiaohui40, liuzhiyuan8, liuchao397, liupengzhang, jiangqixia}@jd.com

Abstract

Representation learning is fundamental to modern machine learning, powering applications such as text retrieval and multimodal understanding. However, learning robust and generalizable representations remains challenging. While prior work has demonstrated that active noise injection, a form of data augmentation, can enhance encoding performance, most existing methods rely on heuristic or static noise, overlooking the dynamic nature of feature distributions during training. In this work, we systematically study the role of noise in representation learning from both gradient-based and feature distribution perspectives, using InfoNCE loss as a representative example. Focusing on multimodal representation learning, we propose **FANoise**, a novel feature-adaptive noise injection strategy. By leveraging the dynamics of contrastive learning, FANoise effectively mitigates the negative impacts of noise while preserving its benefits. Under this theoretically grounded framework, comprehensive experiments demonstrate that FANoise consistently improves overall performance on multimodal tasks across various base VLM models.

Introduction

Representation learning, which aims to capture meaningful and transferable features from raw data, has become a cornerstone of modern machine learning. It plays a pivotal role across diverse applications, from text retrieval (e.g., (Xiao et al. 2023; Li et al. 2023c)) to multimodal understanding (e.g., (Radford et al. 2021; Jia et al. 2021; Jiang et al. 2024b; Wei et al. 2024; Ren et al. 2024; Zhang et al. 2024a; Liu et al. 2022)). Despite its remarkable successes, learning robust and generalizable representations remains challenging.

In multimodal representation learning, contrastive learning frameworks such as CLIP (Radford et al. 2021), ALIGN (Jia et al. 2021), SigLIP (Zhai et al. 2023), and BLIP (Li et al. 2022) have achieved significant progress. Most current multimodal embedding methods primarily focus on architectural innovations (Li et al. 2022; Wei et al. 2024; Ren et al. 2024; Zhang et al. 2024a; Liu et al. 2022) or on enriching training data via data augmentation (Zhang et al. 2024b; Chen et al. 2025; Zhou et al. 2024), such as applying transformations to existing samples or generating new synthetic data. While data augmentation serves to increase data diversity and thereby improve model robustness, it does so by implicitly introducing variations or perturbations at the data level. In contrast, explicit noise injection strategies operate directly at the representation or feature level, offering a more controllable and theoretically analyzable approach to improving robustness and generalization. However, the systematic study of such representation-level noise injection, especially in complex multimodal settings, remains largely unexplored.

Recently, active noise injection approaches such as CLAE (Ho and Nvasconcelos 2020), SimCSE (Gao, Yao, and Chen 2021), and NEFTune (Jain et al. 2023) have demonstrated the effectiveness of controlled noise injection strategies including adversarial noise, dropout masks, and feature perturbations in improving representation quality, particularly in unimodal scenarios. These studies highlight the crucial role of both explicit and implicit noise injection mechanisms in representation learning. However, most existing methods rely on heuristic or static noise augmentation schemes, without explicitly modeling the underlying feature distributions or adapting to the dynamic nature of training. This leads to several fundamental open questions:

- *What are the underlying mechanisms by which noise injection affects representation learning?*
- *How can we design and adapt optimal noise injection strategies for different feature structures and learning scenarios?*
- *How do theoretical principles of noise injection lead to practical improvements in robustness and generalization?*

Addressing these questions is essential for developing a systematic and theoretically grounded framework for multimodal representation learning.

Motivated by these open questions and the limitations of existing approaches, this work systematically investigates the role of noise in multimodal representation learning from both gradient-based and feature distribution perspectives. Specifically, we analyze how noise injection affects the gradient dynamics of the InfoNCE loss and examine its impact on feature distributions through the lens of singular value de-

*These authors contributed equally.

†Corresponding author

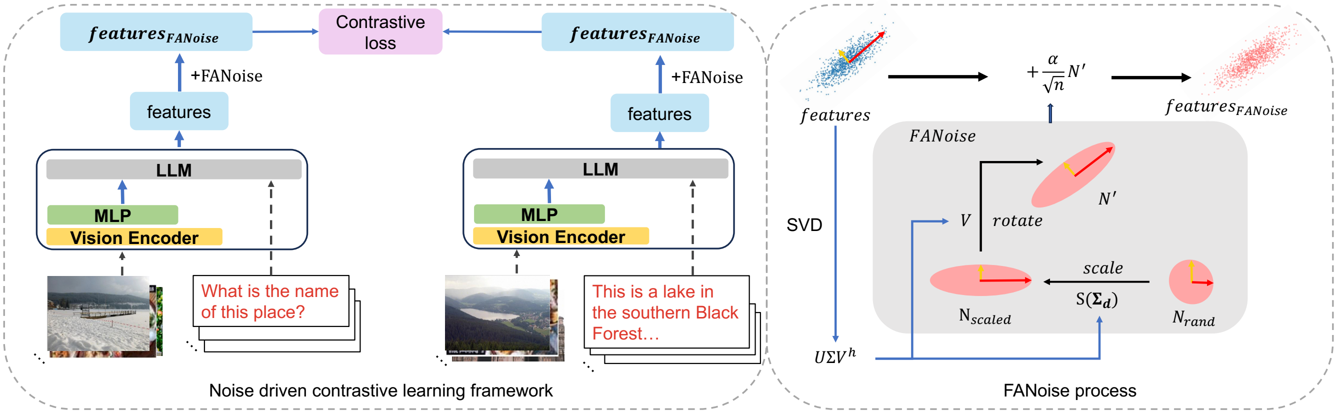

Figure 1: A brief description of our idea. Left: we use a VLM as the backbone to deeply integrate image and text features, both query and target can be image, text or composed image-text data. Before the contrastive objective, we actively inject FANoise into features to improve representation quality. Right: detail of FANoise process, modulate Gaussian noise in V space by singular values after SVD and add it to the original features.

composition (SVD). Based on these theoretical insights, we propose a novel **F**eature-**A**daptive **Noise** injection strategy, termed **FANoise**, which actively injects noise into features before the contrastive objective to improve representation quality. We provide an overview of FANoise in Fig. 1.

Our contributions are threefold:

1. We provide a thorough empirical and theoretical analysis of how noise injection influences feature alignment and uniformity in multimodal contrastive learning, bridging the gap between empirical practice and theoretical understanding.
2. We systematically characterize the optimal noise strength and distribution for effective multimodal representation learning, revealing insights that previous heuristic approaches have overlooked.
3. We propose and validate a feature-adaptive noise augmentation strategy (FANoise) based on singular value decomposition, achieving significant improvements in downstream multimodal tasks.

Through extensive experiments on established multimodal benchmarks, we demonstrate that our approach leads to more robust and generalizable representations, significantly outperforming existing methods. By clarifying the precise role of noise injection in multimodal representation learning, this work provides actionable guidelines for future research and practice.

Related Works

Multimodal Embedding Early research employs dual-encoder models to extract and align features through contrastive loss, including CLIP (Radford et al. 2021), ALIGN (Jia et al. 2021), SigLIP (Zhai et al. 2023), and BLIP (Li et al. 2022). Later work explores feature integration strategies: UniIR (Wei et al. 2024) uses score fusion for separate embeddings, while VISTA (Ren et al. 2024) enhances text encoders with visual fusion modules (Zhang et al. 2024a; Liu et al. 2022). However, these models struggle with deep-level integration, limiting performance in complex retrieval tasks (Zhang et al. 2024b).

Vision-Language Models (VLMs) (Wang et al. 2024; Abdin et al. 2024; Liu et al. 2024) have emerged as popular backbones for multimodal embedding due to their ability to handle diverse image-text combinations and integrate features deeply within transformer architectures. E5-V (Jiang et al. 2024a) and VLM2Vec (Jiang et al. 2024b) utilize instruction tuning to transform VLMs into embedding models. LLaVE (Lan et al. 2025) enhances embeddings through hardness-weighted contrastive learning with reward models, while UniME (Gu et al. 2025) employs a two-stage framework with textual knowledge distillation and hard-negative enhanced tuning.

Contrastive Learning Contrastive learning has achieved remarkable success in modern deep learning through advances in neural architectures and data augmentation techniques. Early unsupervised methods relied on generative approaches like autoencoders (Vincent et al. 2008), which struggled with high-dimensional data. Instance discrimination (Wu et al. 2018) and contrastive objectives (e.g., CPC (Oord, Li, and Vinyals 2018a)) demonstrated that pulling positive pairs together while pushing negatives apart yields transferable features without labels. Key innovations include memory banks (Wu et al. 2018), momentum encoders (He et al. 2020), and temperature-scaled losses (Chen et al. 2020). Subsequent work explores negative-free methods (Grill et al. 2020) and hybrid approaches combining contrastive learning with clustering (Caron et al. 2020) or masked modeling (Bao et al. 2021), extending beyond vision to NLP (Logeswaran and Lee 2018), graphs (Veličković et al. 2018), and multimodal tasks (Radford et al. 2021).

Noise Learning Noise presents both challenges and opportunities in machine learning. Label noise with incorrect annotations is addressed through robust loss functions (Zhang, Gong, and Choi 2021; Wang et al. 2019; Li et al. 2023b), regularization (Shorten and Khoshgoftaar 2019; Szegedy et al. 2016; Liu et al. 2020), label correction (Lu and He 2022; Gong et al. 2022; Zhang et al. 2021), and sample selection (Kim et al. 2019; Han et al. 2018; Xia et al. 2023). Conversely, structured noise in embeddings enhances generalization when controlled. NEFTune (Jain et al. 2023) shows that injecting controlled noise into embeddings during fine-tuning improves LLM robustness. Similarly, stochastic weight averaging (Izmailov et al. 2018) and manifold mixup (Verma et al. 2019) leverage noise to smooth decision boundaries.

In generative modeling, Denoising Diffusion Probabilistic Models (Ho, Jain, and Abbeel 2020) systematically corrupt and reconstruct data, transforming noise into structured samples. For self-supervised learning, noise generates positive pairs for contrastive objectives: SimCSE (Gao, Yao, and Chen 2021) employs dropout noise for augmented views, while MoCo (He et al. 2020) and BYOL (Grill et al. 2020) rely on noise-induced augmentations for invariant representations. Thus, noise is detrimental in supervised settings with label corruption but beneficial in embeddings, contrastive learning, and generative modeling.

Method

In this section, we conduct an in-depth analysis of the noise mechanism and propose a method to effectively incorporate noise with contrastive learning.

Noise Driven Contrastive Learning

InfoNCE Loss InfoNCE Loss (Oord, Li, and Vinyals 2018b) has emerged as the de facto standard in representation learning, serving as a widely used loss function for contrastive learning. We use the standard InfoNCE loss as our training objective:

$$L(\mathbf{q},\mathbf{k}) = -\sum_{l=1}^{N} \log \frac{\exp(\mathbf{q_l}\cdot\mathbf{k_l}/\tau)}{\sum_{i=1}^{N}\exp(\mathbf{q_l}\cdot\mathbf{k_i}/\tau)}. \tag{1}$$

Formally, for a batch of training data $\{(x_1,y_1),...,(x_N,y_N)\}$, the vectors $\mathbf{q_l}$ and $\mathbf{k_l}$ are the normalized embeddings of the x_l and y_l, τ is temperature parameter. In this work, both x_l and y_l can be image, text, or composed image-text data.

The Gradient Of InfoNCE Loss To analyze the influence of gradients on representation, we derive the gradient of InfoNCE loss with respect to the model parameters $\boldsymbol{\theta}$:

$$\nabla_{\boldsymbol{\theta}} L(\mathbf{q},\mathbf{k}) = \sum_{l}^{N} \nabla_{\mathbf{q_l}} L(\mathbf{q},\mathbf{k}) \nabla_{\boldsymbol{\theta}}\mathbf{q_l} + \sum_{i}^{N} \nabla_{\mathbf{k_i}} L(\mathbf{q},\mathbf{k}) \nabla_{\boldsymbol{\theta}}\mathbf{k_i}. \tag{2}$$

Separately, we derive $\nabla_{\mathbf{q_l}} L(\mathbf{q},\mathbf{k})$ and $\nabla_{\mathbf{k_i}} L(\mathbf{q},\mathbf{k})$:

$$\nabla_{\mathbf{q_l}} L(\mathbf{q},\mathbf{k}) = -\mathbf{k_l}/\tau + \overline{\mathbf{k_l}}/\tau, \tag{3a}$$
$$\nabla_{\mathbf{k_i}} L(\mathbf{q},\mathbf{k}) = -\mathbf{q_i}/\tau + \widetilde{\mathbf{q_i}}/\tau. \tag{3b}$$

Here defines $\overline{\mathbf{k_l}} = \sum_{j=0}^{N} p_{lj}\mathbf{k_j}$, $\widetilde{\mathbf{q_j}} = \sum_{l=0}^{N} p_{lj}\mathbf{q_l}$ and $p_{lj} = \exp(\mathbf{q_l}\cdot\mathbf{k_j}/\tau)/\sum_{i=0}^{N}\exp(\mathbf{q_l}\cdot\mathbf{k_i}/\tau)$. Notably, the normalization condition $\sum_{j=0}^{N} p_{lj} = 1$ holds for each l, while no such guarantee holds when summing over l for fixed j.

Finally, we get:

$$\nabla_{\boldsymbol{\theta}} L(\mathbf{q},\mathbf{k}) = -\frac{1}{\tau}\left[\sum_{l}^{N}(\mathbf{k_l}-\overline{\mathbf{k_l}})\nabla_{\boldsymbol{\theta}}\mathbf{q_l} + \sum_{i}^{N}(\mathbf{q_i}-\widetilde{\mathbf{q_i}})\nabla_{\boldsymbol{\theta}}\mathbf{k_i}\right]. \tag{4}$$

The gradient's purpose here is evident - it reduces the distance between $\mathbf{q_i}$ and $\mathbf{k_i}$ while maximizing deviation from the overall weighted mean direction, thus enhancing the isotropy of the representation space.

Noise Driven InfoNCE Loss To demonstrate the function of slight noise, which causes fluctuations in the encoding of $\mathbf{q},\mathbf{k}$, we add Gaussian noise to $\mathbf{k}$ as an example for analysis and assume that the noise to $\mathbf{k_l}$ is $\boldsymbol{\epsilon_l}$. Based on Eq.(3a), the derivative of InfoNCE Loss with respect to the noise is:

$$\nabla_{\mathbf{q_l}} L(\mathbf{q},\mathbf{k}+\boldsymbol{\epsilon}) = -\frac{1}{\tau}(\mathbf{k_l}+\boldsymbol{\epsilon_l}-\overline{(\mathbf{k}+\boldsymbol{\epsilon})_l}). \tag{5}$$

Based on Taylor expansion and expectation of noise distribution, Eq.(5) can be expanded as:

$$\mathbf{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\delta^2)}\left[\nabla_{\mathbf{q_l}} L(\mathbf{q},\mathbf{k}+\boldsymbol{\epsilon})\right] = -\frac{1}{\tau}(\mathbf{k_l}-\overline{\mathbf{k_l}}-\delta^2\mathbf{q_l}/\tau), \tag{6}$$

When the $\mathbf{k}$ sampling distribution is uniform enough, $\mathbf{k_l}$ is parallel to $\mathbf{q_l}$, assuming $\overline{\mathbf{k_l}} \approx \gamma_l\mathbf{q_l}$:

$$\mathbf{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\delta^2)}\left[\nabla_{\mathbf{q_l}} L(\mathbf{q},\mathbf{k}+\boldsymbol{\epsilon})\right] \approx -\frac{1}{\tau}[\mathbf{k_l}-(1+\delta^2/\tau\gamma_l)\overline{\mathbf{k_l}}]. \tag{7}$$

Under the premise of InfoNCE loss, the noise term is equivalent to giving negative samples a larger weight. It also makes $\mathbf{q_l}$ moves away from $\overline{\mathbf{k}}$ and closer to $\mathbf{k_l}$ more quickly.

Significantly, noise on $\mathbf{k}$ contributes no more terms in Eq.(3b) under expectation of noise distribution:

$$\mathbf{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\delta^2)}\left[\nabla_{\mathbf{k_i}} L(\mathbf{q},\mathbf{k}+\boldsymbol{\epsilon})\right] = -\frac{1}{\tau}(\mathbf{q_i}-\widetilde{\mathbf{q_i}}). \tag{8}$$

Feature-Adaptive Noise Distributions

Motivation and Problem Statement Data augmentation through noise injection has proven essential for improving model robustness and generalization. However, conventional approaches apply uniform noise across all feature dimensions, ignoring the inherent heterogeneity in feature importance and signal strength. This uniform treatment leads to two critical issues: (1) the noise energy will increase linearly with the feature dimension, creating instability in high-dimensional spaces, and (2) important features with weak signal strength may be overwhelmed by noise, degrading their signal-to-noise ratio below the discrimination threshold and causing the loss of critical discriminative information. To address these challenges, we propose **FANoise** (Feature-Adaptive Noise injection), which adapts noise intensity according to the spectral characteristics of the data.

Spectral Structure Analysis Understanding the spectral properties of data is fundamental to our approach. Let $\mathbf{X} \in \mathbb{R}^{m\times n}$ denote a data matrix with m samples and n features. Through singular value decomposition (SVD), we obtain its spectral representation:

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^\top, \tag{9}$$

where $\mathbf{U} \in \mathbb{R}^{m\times m}$ and $\mathbf{V} \in \mathbb{R}^{n\times n}$ are unitary matrices, and $\Sigma \in \mathbb{R}^{m\times n}$ contains non-negative singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ ($r = \text{rank}(\mathbf{X})$). The columns of $\mathbf{V}$ define principal directions that capture feature correlations, while singular values quantify the energy concentration in each direction. This decomposition reveals the natural hierarchy of feature importance, with larger singular values corresponding to more significant patterns in the data.

Limitations of Conventional Noise Injection Traditional data perturbation methods apply isotropic Gaussian noise uniformly across all dimensions:

$$\widetilde{\mathbf{X}}_{\text{naive}} = \mathbf{X} + \alpha\mathbf{N}, \tag{10}$$

where $\mathbf{N} \sim \mathcal{N}^{m\times n}(0, \mathbf{I})$ and $\alpha > 0$ controls noise magnitude. This approach suffers from a fundamental dimensional scaling problem. The noise covariance structure $\text{Cov}(\alpha\mathbf{N}) = \alpha^2\mathbf{I}_n$ results in total noise energy $\text{tr}(\alpha^2\mathbf{I}_n) = \alpha^2 n$, which grows linearly with feature dimension n. In high-dimensional spaces, this cumulative noise energy eventually overwhelms the signal, creating a dimensional curse that degrades model performance.

To mitigate this dimensional dependency, a common solution scales the noise injection as:

$$\widetilde{\mathbf{X}} = \mathbf{X} + \frac{\alpha}{\sqrt{n}}\mathbf{N}. \tag{11}$$

The $\frac{1}{\sqrt{n}}$ scaling factor ensures constant total noise energy α^2, independent of dimension n. However, this uniform scaling still fails to account for the heterogeneous importance of different feature directions.

Theoretical Foundation for Adaptive Scaling The need for feature-adaptive noise injection becomes evident when analyzing how noise affects different spectral components. As noise strength increases, dimensions associated with smaller singular values suffer disproportionate signal-to-noise ratio (SNR) degradation. This phenomenon arises from three interconnected theoretical principles:

Spectral Perturbation Theory: According to Weyl's perturbation theorem (Tao 2012), for spectral norm-bounded noise $\mathbf{N}$, the singular values satisfy:

$$|\sigma_i(\widetilde{\mathbf{X}}) - \sigma_i(\mathbf{X})| \leq \sigma_0\left(\frac{\alpha}{\sqrt{n}}\mathbf{N}\right). \tag{12}$$

This bound reveals that dominant features with $\sigma_i(\mathbf{X}) \gg \sigma_0(\frac{\alpha}{\sqrt{n}}\mathbf{N})$ remain stable under noise perturbation, while marginal features with $\sigma_i(\mathbf{X}) \approx \sigma_0(\frac{\alpha}{\sqrt{n}}\mathbf{N})$ risk complete SNR collapse.

Random Matrix Theory: The effective covariance structure of noisy data exhibits a phase transition phenomenon described by spiked covariance models (Perry et al. 2016, 2018):

$$\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}} = \sum_{i=1}^{r}\sigma_i^2\mathbf{v}_i\mathbf{v}_i^\top + \frac{m}{n}\alpha^2\mathbf{I}. \tag{13}$$

This decomposition shows that singular values above the critical threshold $\tau^* = \sqrt{\frac{m}{n}}\alpha$ remain separable from the noise bulk described by the Marchenko-Pastur distribution (Marchenko and Pastur 1967), while those below become statistically indistinguishable from noise.

Information-Theoretic Perspective: From an information preservation standpoint, uniform noise injection treats all feature directions equally, potentially destroying valuable information in weaker but discriminative dimensions. This motivates the need for adaptive scaling that preserves the relative importance hierarchy while maintaining robustness benefits.

Feature-Adaptive Noise Injection Method Building on the theoretical foundation, we propose FANoise, which adapts noise intensity according to local signal strength in each principal direction. The method operates through a two-stage process that systematically addresses the feature importance preservation and dimensional scaling problem:

Stage 1 - Spectral-Aware Noise Generation and Transformation: We first generate noise in the principal component space with adaptive scaling based on local signal strength:

$$\mathbf{N}_{\textbf{scaled}} = \mathbf{N}_{\textbf{rand}} \odot \mathbf{S}(\Sigma_d), \tag{14}$$

where $\mathbf{N}_{\textbf{rand}} \sim \mathcal{N}^{m\times r}(0, 1)$, $\Sigma_d \in \mathbb{R}^r$ contains the diagonal singular values, and $\mathbf{S}(\Sigma_d)$ is the adaptive scaling function that modulates noise intensity based on local signal strength. This adaptive scaling addresses the feature importance preservation problem by ensuring that noise energy is distributed according to the spectral structure.

The scaled noise is then transformed back to the original feature space:

$$\mathbf{N}' = \mathbf{N}_{\textbf{scaled}}\mathbf{V}^\top, \tag{15}$$

where $\mathbf{V}^\top \in \mathbb{R}^{r\times n}$ contains only the effective principal components, ensuring $\mathbf{N}' \in \mathbb{R}^{m\times n}$.

Scaling Function Design We investigate three principled scaling strategies, each with distinct theoretical motivations:

- **Uniform scaling**: $\mathbf{S} = \mathbf{1}$ (equivalent to Eq.(10))
  This serves as our baseline, applying equal noise intensity across all directions.
- **Linear scaling**: $\mathbf{S} = \Sigma_d/\overline{\Sigma_d}$
  This approach scales noise proportionally to signal strength, maintaining constant signal-to-noise ratios and preserving weaker features' discriminative capacity.
- **Sublinear scaling**: $\mathbf{S} = \sqrt{\Sigma_d}/\overline{\sqrt{\Sigma_d}}$
  This strategy provides a compromise between uniform and linear scaling, offering moderate protection for

weaker features while ensuring sufficient perturbation for robustness.

where $\bar{\cdot}$ denotes same as Eq.(3a).

Stage 2 - Dimensionally-Normalized Noise Injection: The final noise injection applies proper dimensional normalization:

$$\widetilde{\mathbf{X}} = \mathbf{X} + \frac{\alpha}{\sqrt{n}}\mathbf{N}', \tag{16}$$

As discussed above, the $\frac{\alpha}{\sqrt{n}}$ ensures that the total noise energy remains constant regardless of the feature dimension n.

Experiments

This section evaluates the effectiveness of FANoise. We begin by introducing the datasets and implementation details, and carry out comparative experiments with state-of-the-art baselines. Finally, we conduct an in-depth analysis of noise strength and distribution.

Dataset Overview and Experiment Setting

Dataset Overview The MMEB (Jiang et al. 2024b) [1] [2] (Massive Multimodal Embedding Benchmark) is a comprehensive benchmark that comprises 36 datasets divided into four meta-tasks: classification, visual question answering, retrieval, and visual grounding. It has 20 in-distribution datasets for training and 36 for testing, with 20 being in-distribution (IND) and 16 out-of-distribution (OOD). We calculate Precision@1 of each dataset, which measures the proportion of top-ranked candidates that are positive samples. MMEB aims to provide a diverse framework for training and evaluating instruction following multimodal embedding models, assessing their generalization across tasks and domains.

Experiment Setting Our implementation builds on the Qwen2-VL-2B (Wang et al. 2024) following the VLM2Vec (Jiang et al. 2024b), and run all experiments on 8 H800 GPUs with a random seed of 42. The adaptation pipeline incorporates LoRA (Hu et al. 2022) adapters (rank=8) for parameter-efficient fine-tuning, processing high-resolution images, each equivalent to 4096 token sequences, with a maximum of 100,000 samples per train dataset. Optimization uses a linear learning rate scheduler (initial rate=2e-5) over 2,000 training steps with 200 warmup iterations, executing on 256 samples per GPU batch. To address memory constraints, GradCache (Gao et al. 2021) is implemented with chunk sizes of 4 for query and candidate embeddings. Critical design choices regarding noise injection parameters are systematically analyzed in Section Result Analysis.

Baselines Following VLM2Vec, we experiment with different backbone models (Qwen2-VL-2B (Wang et al. 2024), LLaVA-NeXT (Liu et al. 2024) and Phi3.5-V-4B (Abdin et al. 2024)). Additionally, we benchmark our results against contemporary advanced models: CLIP (Radford et al. 2021), BLIP2 (Li et al. 2023a), UniIR (Wei et al. 2024), MagicLens (Zhang et al. 2024a), VLM2Vec (Jiang et al. 2024b), UniME (Gu et al. 2025), MegaPairs (Zhou et al. 2024), LLaVE (Lan et al. 2025).

Main Result

In Table 1, we present the performance of our method, where FANoise$_{ss}$ represents *sublinear scaling* noise, achieving the best results among the three distributions as verified in Section Noise Distribution. Following the experimental setup of VLM2Vec (Jiang et al. 2024b), **FANoise**$_{ss}$ achieves an average improvement of 2.04% across five backbones (Phi3.5-V, Qwen2-VL-2B [3], LLaVA-1.6-LR, LLaVA-1.6-HR and Qwen2-VL-7B), with notable improvements of 4.2% for LLaVA-1.6-LR and 3.5% for LLaVA-1.6-HR.

The best-performing model, FANoise$_{ss}$ (LLaVA-1.6-HR), outperforms contemporary SOTAs (UniME (Gu et al. 2025), MegaPairs (Zhou et al. 2024), VLM2Vec (Jiang et al. 2024b)) by 0.6% over the best-performing Qwen2-VL-7B [4]. Our UniME (Gu et al. 2025) results are from re-evaluating their checkpoint, as their original 'average overall score' wasn't a simple average across all 36 datasets.

LLaVA-OV-7B serves as a strong baseline, achieving comparable performance to Qwen2-VL-7B (65.8%) in VLM2Vec (Lan et al. 2025). LLaVE (Lan et al. 2025) further optimizes it to reach 70.3%. Due to GPU constraints, we have not extended experiments to LLaVA-OV-7B (Li et al. 2024).

Both LLaVE and UniME enhance discriminative power by prioritizing hard-negative samples during training; LLaVE uses a reward model for higher weights, while UniME filters false negatives and samples hard negatives. As shown by Eq.(7), FANoise effectively weights negative samples by noise, aligning with this approach. Additionally, it excels in simplicity, effectiveness, and integration ease:

1. Generalized Robustness Across Architectures: FANoise enhances representation robustness and generalizes well to various model architectures and capacities.
2. Seamless Integration: Its plug-and-play design integrates seamlessly into existing training pipelines without redesign costs.

Result Analysis

To enable flexible and efficient result analysis, we use a batch size of 16 per GPU on 8 GPUs with up to 5,000 training samples per dataset, employing Qwen2-VL-2B (Wang et al. 2024) as the backbone model.

First, we examine the impact of noise strength in these settings, identifying an optimal range that balances robustness and precision. Second, we compare various noise distributions, demonstrating that our proposed feature-adaptive noise consistently outperforms uniform Gaussian noise. These experiments collectively confirm the effectiveness of our method in critical design choices.

Noise Strength Based on FANoise$_{ss}$, we examine the impact of noise strength (α) on model performance, as illustrated in Fig. 2. All experiments show improvements over

[1] https://huggingface.co/datasets/TIGER-Lab/MMEB-train
[2] https://huggingface.co/datasets/TIGER-Lab/MMEB-eval
[3] https://huggingface.co/TIGER-Lab/VLM2Vec-Qwen2VL-2B
[4] https://huggingface.co/TIGER-Lab/VLM2Vec-Qwen2VL-7B

Model	Per Meta-Task Score				Avg Score		
	Classification	VQA	Retrieval	Grounding	IND	OOD	Overall
# Datasets	10	10	12	4	20	16	36
Zero-shot on MMEB							
CLIP	42.8	9.1	53.0	51.8	37.1	38.7	37.8
BLIP2	27.0	4.2	33.9	47.0	25.3	25.1	25.2
UniIR(BLIP$_{FF}$)	42.1	15.0	60.1	62.2	44.7	40.4	42.8
UniIR(CLIP$_{SF}$)	44.3	16.2	61.8	65.3	47.1	41.7	44.7
Magiclens	38.8	8.3	35.4	26.0	31.0	23.7	27.8
Fine-tuning on MMEB							
UniME(Phi3.5-V)	54.6	55.9	64.5	81.8	68.2	52.7	61.3
UniME(LLaVA-1.6)	60.6	52.9	67.9	85.1	68.4	57.9	63.6
MegaPairs(LLaVA-1.6)	56.0	57.4	69.9	83.6	68.0	59.1	64.1
VLM2Vec(LLaVA-OV-7B)	63.5	61.1	64.5	87.3	69.7	61.0	65.8
LLaVE(LLaVA-OV-7B)	65.7	65.4	70.9	91.9	75.0	64.4	70.3
VLM2Vec(Phi3.5-V)	54.8	54.9	62.3	79.5	66.5	52.0	60.1
FANoise$_{ss}$(Phi3.5-V)	54.6	**56.1**	**63.9**	78.6	**68.2**	51.5	**60.8(+0.7)**
VLM2Vec(Qwen2-VL-2B)	59.0	49.4	65.4	73.4	66.0	52.6	60.1
FANoise$_{ss}$(Qwen2-VL-2B)	**60.3**	**51.3**	**66.2**	72.3	**67.2**	**53.4**	**61.1(+1.0)**
VLM2Vec(LLaVA-1.6-LR)	54.7	50.3	56.2	64.0	61.0	47.5	55.0
FANoise$_{ss}$(LLaVA-1.6-LR)	**57.0**	43.1	**64.1**	**90.1**	**66.1**	**50.5**	**59.2(+4.2)**
VLM2Vec(LLaVA-1.6-HR)	61.2	49.9	67.4	86.1	67.5	57.1	62.9
FANoise$_{ss}$(LLaVA-1.6-HR)	**62.2**	**57.3**	**69.1**	**91.4**	**72.2**	**59.1**	**66.4(+3.5)**
VLM2Vec(Qwen2-VL-7B)	62.6	57.8	69.9	81.7	72.2	57.8	65.8
FANoise$_{ss}$(Qwen2-VL-7B)	**63.7**	**59.4**	**70.3**	80.4	**73.5**	**57.9**	**66.6(+0.8)**

Table 1: Performance comparison on the MMEB benchmark. The FF and SF subscripts under CLIP or BLIP represent feature-level fusion and score-level fusion, respectively. LR suffixes signify training and inference on low-resolution (336×336) images, while HR suffixes similarly denote both processes on high-resolution (1344×1344) images. Reported scores are the average Precision@1 over the corresponding datasets.

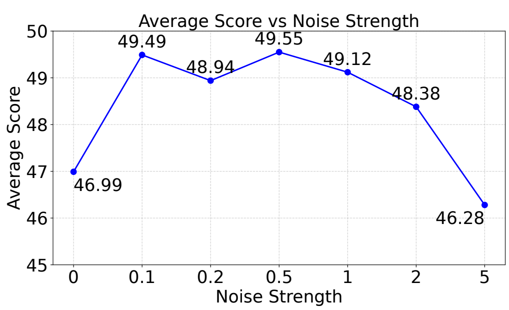

Figure 2: Performance comparison of different noise strengths on the MMEB benchmark

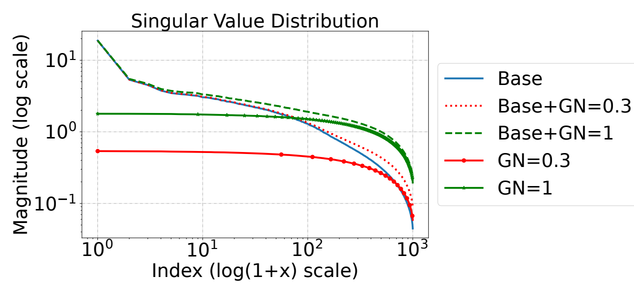

Figure 3: Log-log Plot of Singular Values: Base Embedding, Gaussian Noise(GN), and Noisy Embedding

the no-noise baseline except at overly high noise strength, indicating that excessive noise destroys tail features. Performance initially improves with increasing α before declining.

To explore the interaction between noise strength and features, we disturb base embeddings with uniform Gaussian noise controlled by $\alpha/\sqrt{n}$, where α is set to **0.3** and **1**. Using 1000 randomly sampled test data points with feature dimension $n = 1536$, we perform Singular Value Decomposition (SVD) on three matrices:

- Base feature matrix(F): Extracted from test data points;
- Pure Gaussian noise matrix(GN): Same dimension as embeddings, with entries from $\mathcal{N}(0, (\alpha/\sqrt{n})^2)$;
- Noisy embedding matrix($F+GN$): Generated by adding scaled noise to base embeddings.

The singular values of these matrices are plotted on a log-log scale, with the x-axis representing the index of singular values and the y-axis their magnitudes.

Spectral Perturbation Analysis As observed in Fig. 3, the singular values of GN match the theoretical Marchenko-Pastur noise bulk, with boundary $[\left|1-\sqrt{\frac{m}{n}}\right|, 1+\sqrt{\frac{m}{n}}] *$

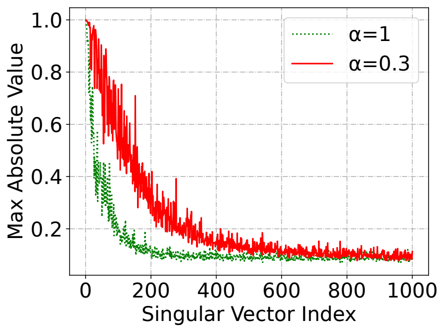

Figure 4: inner product between singular vector of F and $F + GN$

Model	Avg Score
Baseline	60.06
Baseline + Uniform scaling	60.93(+0.87)
Baseline + Linear scaling	60.69(+0.63)
Baseline + Sublinear scaling	**61.08(+1.02)**

Table 2: Three Possible Noise Distribution Results on the MMEB benchmark

$\alpha \approx [0.2, 1.8] * \alpha$. According to Eq. 12, singular values of F greater than the upper boundary $(1.8 * \alpha)$ effectively *escape* the Marchenko-Pastur noise bulk, preserving their signal-to-noise ratio. When singular values fall below this boundary, F experiences significant perturbations and enters a noise-dominated regime. This demonstrates how Weyl's perturbation theorem predicts the stability of dominant features under noise perturbation, while marginal features with comparable singular values to the noise spectral norm risk complete SNR collapse.

Phase Transition Analysis Additionally, Eq. 13 provides a more precise critical threshold: $\tau^* = \sqrt{\frac{m}{n}}\alpha \approx 0.8\alpha$. This threshold marks the boundary between distinguishable (signal-dominated) and indistinguishable (noise-dominated) singular directions. For $\alpha = 0.3/1$, the singular value around $0.24/0.8$ corresponds to index $600/180$. As shown in Fig. 4, starting from the $600/180$-th singular value, the inner products between corresponding singular vectors of (F) and $(F + GN)$ remain almost identical, indicating these singular directions become indistinguishable from noise.

These conclusions align closely with theoretical predictions, clearly elucidating how noise strength impacts the structural integrity of embedding features. In summary, We select $\alpha = 0.1$ for our main experiments, as it consistently achieves competitive performance across various small-batch trials. This value represents a balance between regularization and information disruption, providing sufficient perturbation to prevent overfitting while preserving meaningful signal.

Noise Distribution Because noise distribution matters, we ran full experiments on three noise distributions $\mathbf{S}(\Sigma_d)$ of Eq.(14), each with a fixed noise strength of 0.1. As shown in

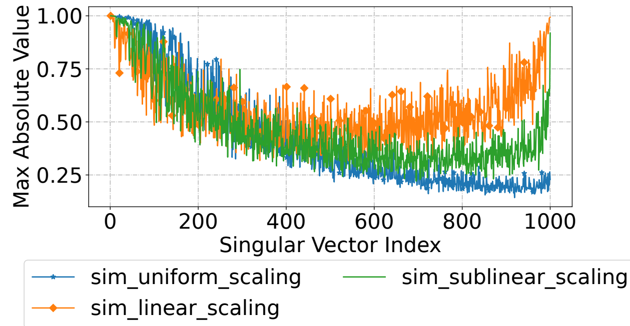

Figure 5: Max Similarity Values: Base Embedding vs. Noisy Embeddings

Fig. 2, the baseline VLM2Vec [5] scores 60.06% on MMEB. Uniform Gaussian noise (*Uniform scaling*) improves performance by 0.87% (to 60.93%), suggesting that appropriate noise enhances model robustness by preventing overfitting. *Linear scaling* shows minimal improvement, showing that omitting the square-root step produces suboptimal noise scaling. FANoise's *Sublinear scaling* achieves the highest score of 61.08%, outperforming the baseline by 1.02%. Under isotropic Gaussian noise (*Uniform scaling*), dimensions with smaller singular values undergo disproportionately larger signal-to-noise ratio (SNR) drops due to weaker signal strength. FANoise protects discriminative features prone to SNR loss.

To analyze why *sublinear scaling* performs best, we compute right-singular vectors $(\mathbf{V}^\top)$ via SVD for base and three noise-augmented embeddings, comparing their similarity to the base embedding (see Fig. 5). *Uniform scaling* shows limited perturbation in higher singular value regions, while *linear scaling* maintains high similarity across lower singular value regions. *Sublinear scaling* demonstrates a different pattern, with moderate perturbation in higher singular value regions and reduced similarity in lower singular value regions, resulting in superior performance.

Conclusion

While most existing noise injection methods rely on heuristic or static approaches that fail to account for the dynamic nature of feature distributions during training, we conduct a systematic investigation from both gradient-based and feature distribution perspectives. Focusing on multimodal representation learning, we propose FANoise - a feature-adaptive noise injection strategy that dynamically responds to evolving feature structures. Our theoretical framework and extensive experiments demonstrate that FANoise consistently enhances robustness and generalization across models of varying architectures and capacities. Designed as a plug-and-play module, FANoise provides an architecture-agnostic solution that complements existing representation learning techniques, offering new possibilities for developing more adaptive and robust learning systems.

[5] https://huggingface.co/TIGER-Lab/VLM2Vec-Qwen2VL-2B

References

Abdin, M.; Aneja, J.; Awadalla, H.; Awadallah, A.; Awan, A. A.; Bach, N.; Bahree, A.; Bakhtiari, A.; Bao, J.; Behl, H.; et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Bao, H.; Dong, L.; Piao, S.; and Wei, F. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33: 9912–9924.

Chen, H.; Wang, L.; Yang, N.; Zhu, Y.; Zhao, Z.; Wei, F.; and Dou, Z. 2025. mmE5: Improving Multimodal Multilingual Embeddings via High-quality Synthetic Data. *arXiv preprint arXiv:2502.08468*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PmLR.

Gao, L.; Zhang, Y.; Han, J.; and Callan, J. 2021. Scaling deep contrastive learning batch size under memory limited setup. *arXiv preprint arXiv:2101.06983*.

Gao, T.; Yao, X.; and Chen, D. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Gong, C.; Ding, Y.; Han, B.; Niu, G.; Yang, J.; You, J.; Tao, D.; and Sugiyama, M. 2022. Class-wise denoising for robust learning under label noise. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 2835–2848.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33: 21271–21284.

Gu, T.; Yang, K.; Feng, Z.; Wang, X.; Zhang, Y.; Long, D.; Chen, Y.; Cai, W.; and Deng, J. 2025. Breaking the Modality Barrier: Universal Embedding Learning with Multimodal LLMs. *arXiv preprint arXiv:2504.17432*.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.; and Sugiyama, M. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

Ho, C.-H.; and Nvasconcelos, N. 2020. Contrastive learning with adversarial examples. *Advances in Neural Information Processing Systems*, 33: 17081–17093.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Izmailov, P.; Podoprikhin, D.; Garipov, T.; Vetrov, D.; and Wilson, A. G. 2018. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*.

Jain, N.; Chiang, P.-y.; Wen, Y.; Kirchenbauer, J.; Chu, H.; Somepalli, G.; Bartoldson, B.; Kailkhura, B.; Schwarzschild, A.; and Saha, A. 2023. Neftune: Noisy Embeddings Improve Instruction Finetuning. arXiv, abs/2310.05914.

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

Jiang, T.; Song, M.; Zhang, Z.; Huang, H.; Deng, W.; Sun, F.; Zhang, Q.; Wang, D.; and Zhuang, F. 2024a. E5-v: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*.

Jiang, Z.; Meng, R.; Yang, X.; Yavuz, S.; Zhou, Y.; and Chen, W. 2024b. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*.

Kim, Y.; Yim, J.; Yun, J.; and Kim, J. 2019. Nlnl: Negative learning for noisy labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, 101–110.

Lan, Z.; Niu, L.; Meng, F.; Zhou, J.; and Su, J. 2025. LLaVE: Large Language and Vision Embedding Models with Hardness-Weighted Contrastive Learning. *arXiv preprint arXiv:2503.04812*.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

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

Li, X. C.; Xia, X.; Zhu, F.; Liu, T.; Zhang, X. Y.; and Liu, C. L. 2023b. Dynamics-aware loss for learning with label noise. *Pattern Recognition*, 144(000): 11.

Li, Z.; Zhang, X.; Zhang, Y.; Long, D.; Xie, P.; and Zhang, M. 2023c. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024. Llavanext: Improved reasoning, ocr, and world knowledge.

Liu, S.; Niles-Weed, J.; Razavian, N.; and Fernandez-Granda, C. 2020. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33: 20331–20342.

Liu, Z.; Xiong, C.; Lv, Y.; Liu, Z.; and Yu, G. 2022. Universal vision-language dense retrieval: Learning a unified representation space for multi-modal retrieval. *arXiv preprint arXiv:2209.00179*.

Logeswaran, L.; and Lee, H. 2018. An efficient framework for learning sentence representations. *arXiv preprint arXiv:1803.02893*.

Lu, Y.; and He, W. 2022. SELC: self-ensemble label correction improves learning with noisy labels. *arXiv preprint arXiv:2205.01156*.

Marchenko, V. A.; and Pastur, L. A. 1967. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018a. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018b. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Perry, A.; Wein, A. S.; Bandeira, A. S.; and Moitra, A. 2016. Optimality and sub-optimality of PCA for spiked random matrices and synchronization. *arXiv preprint arXiv:1609.05573*.

Perry, A.; Wein, A. S.; Bandeira, A. S.; and Moitra, A. 2018. Optimality and sub-optimality of PCA I: Spiked random matrix models. *The Annals of Statistics*, 46(5): 2416–2451.

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

Ren, W.; Yang, H.; Min, J.; Wei, C.; and Chen, W. 2024. VISTA: Enhancing Long-Duration and High-Resolution Video Understanding by Video Spatiotemporal Augmentation. *arXiv preprint arXiv:2412.00927*.

Shorten, C.; and Khoshgoftaar, T. M. 2019. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1): 1–48.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Tao, T. 2012. *Topics in random matrix theory*, volume 132. American Mathematical Soc.

Veličković, P.; Fedus, W.; Hamilton, W. L.; Liò, P.; Bengio, Y.; and Hjelm, R. D. 2018. Deep graph infomax. *arXiv preprint arXiv:1809.10341*.

Verma, V.; Lamb, A.; Beckham, C.; Najafi, A.; Mitliagkas, I.; Lopez-Paz, D.; and Bengio, Y. 2019. Manifold mixup: Better representations by interpolating hidden states. In *International conference on machine learning*, 6438–6447. PMLR.

Vincent, P.; Larochelle, H.; Bengio, Y.; and Manzagol, P.-A. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, 1096–1103.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, Y.; Ma, X.; Chen, Z.; Luo, Y.; Yi, J.; and Bailey, J. 2019. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, 322–330.

Wei, C.; Chen, Y.; Chen, H.; Hu, H.; Zhang, G.; Fu, J.; Ritter, A.; and Chen, W. 2024. Uniir: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision*, 387–404. Springer.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3733–3742.

Xia, X.; Han, B.; Zhan, Y.; Yu, J.; Gong, M.; Gong, C.; and Liu, T. 2023. Combating noisy labels with sample selection by mining high-discrepancy examples. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1833–1843.

Xiao, S.; Liu, Z.; Zhang, P.; and Muennighoff, N. 2023. C-Pack: Packaged Resources To Advance General Chinese Embedding. arXiv:2309.07597.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11975–11986.

Zhang, K.; Luan, Y.; Hu, H.; Lee, K.; Qiao, S.; Chen, W.; Su, Y.; and Chang, M.-W. 2024a. Magiclens: Self-supervised image retrieval with open-ended instructions. *arXiv preprint arXiv:2403.19651*.

Zhang, S.; Gong, C.; and Choi, E. 2021. Learning with Different Amounts of Annotation: From Zero to Many Labels. In *Empirical Methods in Natural Language Processing*.

Zhang, X.; Zhang, Y.; Xie, W.; Li, M.; Dai, Z.; Long, D.; Xie, P.; Zhang, M.; Li, W.; and Zhang, M. 2024b. GME: Improving Universal Multimodal Retrieval by Multimodal LLMs. *arXiv preprint arXiv:2412.16855*.

Zhang, Y.; Zheng, S.; Wu, P.; Goswami, M.; and Chen, C. 2021. Learning with feature-dependent label noise: A progressive approach. *arXiv preprint arXiv:2103.07756*.

Zhou, J.; Liu, Z.; Liu, Z.; Xiao, S.; Wang, Y.; Zhao, B.; Zhang, C. J.; Lian, D.; and Xiong, Y. 2024. MegaPairs: Massive Data Synthesis For Universal Multimodal Retrieval. *arXiv preprint arXiv:2412.14475*.