Neighbor Does Matter: Curriculum Global Positive-Negative Sampling for Vision-Language Pre-training

Bin Huang*†
huangb23@mails.tsinghua.edu.cn
DCST, Tsinghua University
Beijing, China

Feng He*
252135402@qq.com
Baidu Inc.
Beijing, China

Qi Wang
wangqi31@baidu.com
Baidu Inc.
Beijing, China

Hong Chen
h-chen20@mails.tsinghua.edu.cn
DCST, Tsinghua University
Beijing, China

Guohao Li
liguohao@baidu.com
Baidu Inc.
Beijing, China

Zhifan Feng
fengzhifan@baidu.com
Baidu Inc.
Beijing, China

Xin Wang‡
xin_wang@tsinghua.edu.cn
DCST, BNRist, Tsinghua University
Beijing, China

Wenwu Zhu‡
wwzhu@tsinghua.edu.cn
DCST, BNRist, Tsinghua University
Beijing, China

Abstract

Sampling strategies have been widely adopted in Vision-Language Pre-training (VLP) and have achieved great success recently. However, the sampling strategies adopted by current VLP works are limited in two ways: i) they only focus on negative sampling, ignoring the importance of more informative positive samples; ii) their sampling strategies are conducted in the local in-batch level, which may lead to sub-optimal results. To tackle these problems, in this paper, we propose a curriculum-based Global Positive-Negative Sampling (GPN-S) framework for vision-language pre-training, which conducts both positive and negative sampling in the global level, grounded on the notion of neighborhood relationships. Additionally, we incorporate curriculum learning into our sampling strategy, progressively increasing the complexity of samples as the training progresses. Specifically, our proposed GPN-S framework is capable of utilizing positive sampling to bring semantically equivalent samples closer, as well as employing negative sampling to push challenging negative samples farther away. We jointly consider them for vision-language pre-training on the global-level perspective rather than a local-level mini-batch, which provides more informative and diverse samples. We evaluate the effectiveness of the proposed GPN-S framework by conducting experiments on several common downstream tasks, and the results demonstrate significant performance improvement over the existing models.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**.

Keywords

Vision-Language Pre-training, Positive Sampling, Negative Sampling, Curriculum Learning

ACM Reference Format:
Bin Huang, Feng He, Qi Wang, Hong Chen, Guohao Li, Zhifan Feng, Xin Wang, and Wenwu Zhu. 2024. Neighbor Does Matter: Curriculum Global Positive-Negative Sampling for Vision-Language Pre-training. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28-November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3664647.3681502

1 Introduction

In recent years, there has been a burgeoning interest in the field of vision-language pre-training (VLP) within the AI community[2, 12, 25, 33, 44]. VLP aims to learn multimodal representations from large-scale image-text pairs that can simultaneously process visual and textual data, with the goal of improving performance on downstream tasks such as image-text retrieval and visual question answering. [1, 15, 23, 24, 27, 28]

Multi-modal alignment is fundamental to the efficacy of VLP, where semantically similar samples are mapped together to enhance representation accuracy. Recent studies[25, 33] have shown that smart sampling strategies can significantly boost VLP model performance. For example, CLIP[33] benefits from using large mini-batch sizes to find hard negatives, while ALBEF[25] uses a large negative queue to identify challenging examples, both leading to notable performance gains. However, two limitations persist:

- Current works predominantly focus on sampling hard negatives within mini-batches[2, 25, 44], a local-level approach that yields less optimal negative samples.
- There is limited research on positive sampling strategies, with many studies overlooking the potential of positive sampling in enriching multi-modal alignment information.

*Equal contribution.
†Work done as an intern in Baidu Inc.
‡Corresponding authors. DCST is the abbreviation of Department of Computer Science and Technology. BNRist is the abbreviation of Beijing National Research Center for Information Science and Technology.

MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3681502

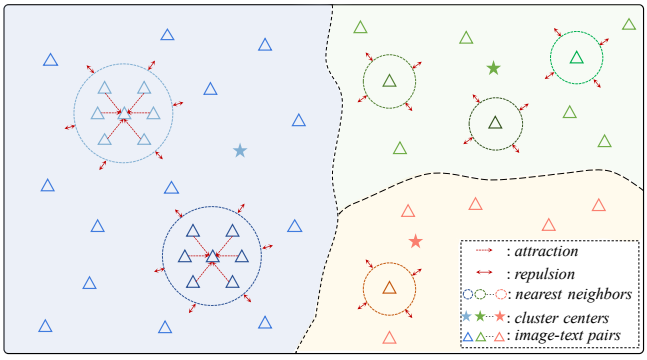

Figure 1: Our proposed sampling strategy involves leveraging information from nearest neighbors to sample semantically equivalent positives, as well as employing a cluster algorithm to obtain global-level challenging negatives.

To overcome these limitations, we propose the curriculum-based Global Positive-Negative Sampling (GPN-S) framework for VLP. This innovative framework conducts both positive and negative sampling on a global level, moving beyond the constraints of local-level mini-batches. Specifically, GPN-S maps the entire dataset's samples into a unified embedding space and utilizes their neighborhood relationships to determine which samples should be closer or farther apart, as illustrated in Figure 1. This approach provides more informative and diverse samples. For global positive sampling (GP-S), we aggregate co-occurring nouns from neighboring texts, replace noisy web-crawled text to enhance the model's cross-modal alignment ability. By incorporating curriculum learning[4, 39, 41, 47, 48], we progressively introduces more complex and challenging samples throughout the training process[6, 7, 21, 22, 32, 45, 49, 50]. We also mine semantically equivalent images to improve the model's uni-modal alignment ability. For global negative sampling (GN-S), we employ a global cluster technique to obtain more challenging negative samples. Notably, our method is model-agnostic and can be applied to enhance most existing VLP models.

In brief, our contributions can be summarized as follows:

- We propose the GPN-S framework, a novel approach that broadens the sampling strategy to encompass both positive and negative pairs at a global level, thus significantly enhancing VLP model performance.
- We utilize the relationships between neighboring samples to effectively integrate information from the embedding space, improving the quality of representations in both cross-modal and uni-modal scenarios.
- We conduct extensive experiments on several downstream tasks to demonstrate that our GPN-S can significantly yield better performance than the models trained without GPN-S.

2 Related Works

2.1 Vision and Language Pre-training

Vision-Language Pre-training (VLP) models are primarily categorized into two types: dual encoder models and fusion encoder models. Dual encoder models, such as CLIP[33] and ALIGN[16], independently encode images and texts, employing contrastive learning to align image-text pair embeddings. While effective in retrieval tasks, their limited interaction between modalities restricts performance in complex tasks like Visual Question Answering (VQA).

Fusion encoder models overcome this by integrating a fusion encoder to meld features from uni-modal encoders. Early iterations[10, 26, 29, 37] relied on pre-trained object detectors for visual feature extraction, facing significant time overhead and capacity constraints. ViLT[19] addressed this by directly inputting image patch features and text token embeddings into a ViT[11] model, but this approach lagged behind object-detector based models due to inadequate uni-modal modeling. Subsequent developments, such as ALBEF[25], combined the benefits of dual and fusion encoders, inspiring further enhancements in models like TCL[44]. TCL introduced intra-modal and global-local contrastive losses to achieve better alignment capabilities.

Our framework distinguishes itself by being model-agnostic, compatible with a broad range of VLP models. It uniquely enhances performance by refining the sampling strategies of these models, a critical factor not explicitly addressed in previous approaches.

2.2 Sampling for Positive and Negative

The primary focus of representation learning lies in extracting semantic information from data, where the representations of similar (positive) pairs are clustered together and those of dissimilar (negative) pairs are spread apart[3]. The sampling of positive and negative pairs is critical in the success of VLP alignment.

Positive sampling traditionally focuses on generating semantically similar pairs, particularly through random data augmentation in uni-modal contexts[8, 9, 38]. Approaches like NNCLR[13] have leveraged the nearest-neighbor sample as the positive. It leverages a queue for nearest-neighbor identification in uni-modal data, which is notably trained on the 1000-class ImageNet dataset. In such a clean, categorized environment, a queue suffices for identifying similar neighbors. However, challenges emerge in web-crawled, noisier datasets where identifying semantically similar positive samples requires more nuanced strategies. This underscores the need for advanced positive sampling approaches in cross-modal scenarios, a domain less explored in previous research.

Negative sampling is aimed at identifying pairs with dissimilar semantics yet closely embedded representations. Techniques that incorporate hard negative sampling have shown to be beneficial. For instance, ALBEF[25] selects the nearest sample in a mini-batch as the hard negative. VLMo[2] extends this approach by mining hard negatives from a broader pool, gathering training examples across all GPUs, In contrast, GRIT-VLP[5] maintains a queue to identify hard negatives, resulting in more significant improvements.

Our work introduces a novel unified positive-negative sampling framework, distinct for its global-level approach. Unlike existing methods constrained by the limited scope of mini-batches or queues, our framework performs sampling across the entire dataset. This comprehensive strategy enables more effective differentiation and identification of positive and negative samples, especially in noisy, diverse datasets, addressing gaps left by previous VLP research and offering a significant advancement in the field.

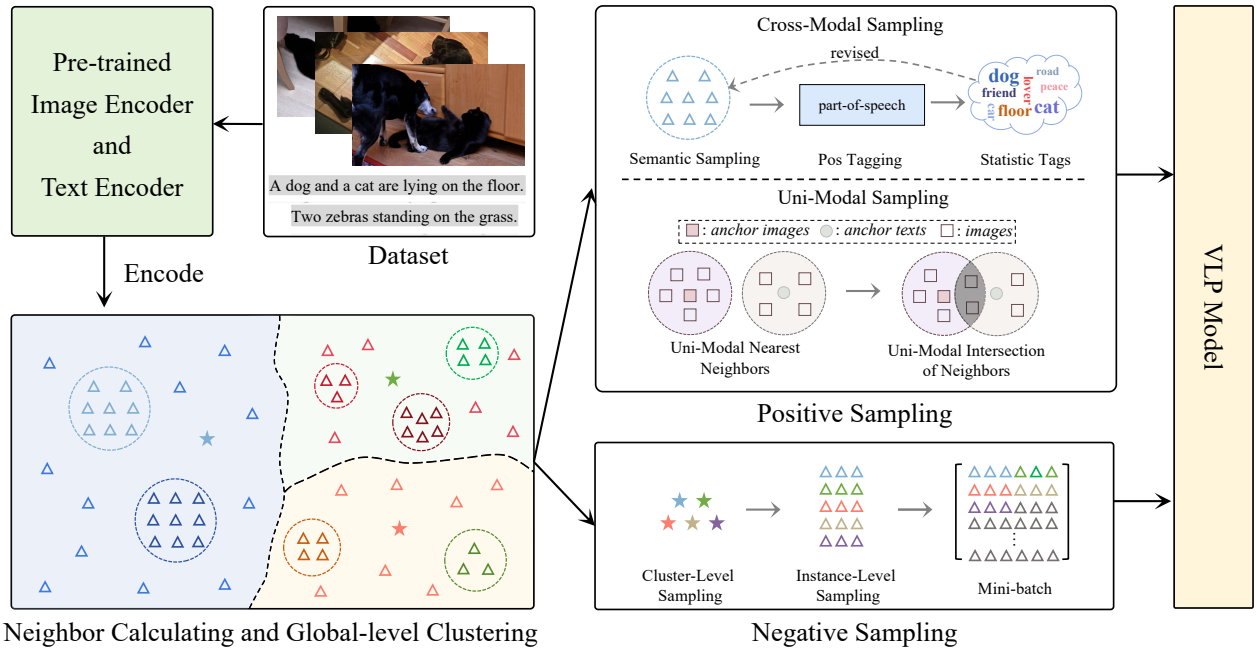

Figure 2: The proposed method contains two primary strategies: 1) Global Positive Sampling(GP-S), which aims to sample both cross-modal and uni-modal data pairs that share same semantic through the neighboring information. 2) Global Negative Sampling(GN-S), which aims to sample more challenging negative data through the clustering information.

3 Method

In this section, we present the Global Positive-Negative Sampling (GPN-S) framework, which is designed to improve the performance of existing VLP models by refining their sampling strategies. We will first introduce the preliminaries about how current VLP models align image and text.

3.1 Preliminaries

For a VLP model composed of a visual encoder $f_V(\cdot)$, a text encoder $f_T(\cdot)$, and a fusion encoder $f_F(\cdot,\cdot)$, the unimodal representations for an image-text pair (v,t) are obtained as follows:

$$\{\mathbf{v}^{\text{cls}}, \mathbf{v}^1, \mathbf{v}^2, ...\} = f_V(v), \tag{1}$$

$$\{\mathbf{t}^{\text{cls}}, \mathbf{t}^1, \mathbf{t}^2, ...\} = f_T(t), \tag{2}$$

where $\mathbf{v}^{\text{cls}}$ and $\mathbf{t}^{\text{cls}}$ represent the embeddings of the corresponding [CLS] token. The fusion representation is obtained through $f_F(f_V(v), f_T(t))$.

Training the VLP model typically involves adopting the image-text matching objective, which introduces a predict head $f_H(\cdot)$ to predict the matching likelihood between the image and text. The probability of a match between v and t is calculated as:

$$p(v,t) = f_H(f_F(f_V(v), f_T(t))). \tag{3}$$

Given a batch of N image-text pairs $\{(v_i,t_i)\}_{i=1}^N$, where (v_i,t_i) is referred to as a positive pair and $(v_i,t_j), \forall j \neq i$ as a negative pair. The image-text matching loss on the positive pair (v_i,t_i) is calculated using binary cross-entropy loss:

$$\mathcal{L}_{\text{itm}}(v_i,t_i) = -log(p(v_i,t_i)) - log(1-p(v_i,t_j)), \tag{4}$$

where $t_j (j \neq i)$ is a randomly selected text from the batch. Recent works[2, 25] introduce hard negative sampling by selecting a t_j to maximize the dot product $(\mathbf{v}_i^{\text{cls}})^T \mathbf{t}_j^{\text{cls}}$ within the batch. The hard negative pairs share similar semantics but differ in fine-grained details, thereby enhancing the model's ability for fine-grained understanding. However, the quality of hard negatives of existing works is constrained by the batch size N.

Different VLP models may introduce different loss functions, all of which are designed to align positive image-text pair (v_i,t_i). Without loss of generality, we denote all other losses as $\mathcal{L}_{\text{other}}$, and the final training loss is expressed as:

$$\mathcal{L}_f(v_i,t_i) = \mathcal{L}_{\text{itm}}(v_i,t_i) + \mathcal{L}_{\text{other}}(v_i,t_i). \tag{5}$$

While current VLP models exhibit outstanding performance across diverse tasks, there are still some limitations in their sampling strategies for positive and negative pairs. Firstly, due to the prohibitive cost of manual labeling, researchers have to train VLP models with web-collected image-text pairs[30, 34]. However, these texts may contain noise and fail to precisely describe the content of corresponding images (e.g., an image of a dog paired with the text *"The photo is taken on Sunday"*), and such positive pairs hinder the alignment process. Additionally, the current sampling is confined to local-level batches with a limited number of samples, thereby restricting the ability to sample optimal positives or negatives.

Figure 3: Illustration of GP-S for cross-modal. We revise the original text based on nouns that frequently appear in neighboring texts.

To address these problems, we propose the Global Positive-Negative Sampling (GPN-S) framework. As shown in Figure 2, GPN-S utilizes a pre-trained encoder, such as CLIP[33], to map image and text data from the entire dataset into a unified global embedding space. Within this space, GPN-S samples optimal positives and negatives by leveraging the neighboring relationships among samples. We will introduce how to calculate the neighboring relationships in section 3.2. Then, in section 3.3 and section 3.4, we respectively give the detailed description for how GPN-S samples positives and negatives through neighboring relationships.

3.2 Neighbor Calculating

GPN-S utilizes an off-the-shelf pre-trained model that consists of a visual encoder $g_V(\cdot)$ and a text encoder $g_T(\cdot)$. These encoders are employed to map image and text from the entire dataset into a unified global embedding space. Given a dataset with D image-text pairs $\{(v_i, t_i)\}_{i=1}^{D}$, it computes the image embedding $\mathbf{v}_i^g = g_V(v_i)$ and text embedding $\mathbf{t}_i^g = g_T(t_i)$ for each pair. Then, we employ Faiss [17] to identify the k nearest-neighbor texts for each image v_i, determined by the closest cosine distance to v_i's embedding, denoted as $V2T_k(v_i)$:

$$V2T_k(v_i) = \{t_j \mid j \in \underset{j' \in [1,D]}{\text{top}_k} \{(\mathbf{v}_i^g)^T \mathbf{t}_{j'}^g\}\}, \tag{6}$$

where top_k identifies the indices of the top k largest values. Similarly, we compute k nearest-neighbor images $V2V_k(v_i)$ for each image v_i, and k nearest-neighbor images $T2V_k(t_i)$ for each text t_i. These cross-modal and uni-modal neighbors are crucial for guiding our sampling strategy.

3.3 Global Positive Sampling (GP-S)

Due to textual noise, the pair (v_i, t_i) does not always serve as a genuinely positive pair that conducive to alignment. We address this issue from two perspectives. From cross-modal perspective, we construct a refined text aligned with v_i. From uni-modal perspective, we establish uni-modal alignment among sampled positive image pairs, enabling the representation of v_i to align with semantically similar text through other images. This section will elaborate on the utilization of global neighboring information for sampling positive cross-modal and uni-modal pairs.

3.3.1 GP-S for Cross-Modal. As mentioned earlier, positive sampling for cross-modal refers to constructing a refined text. We need to extract the most crucial visual information from the image, which pertains to the objects present in the image. We can infer the objects through the neighboring text, as shown in Figure 3. To be specific, we utilize SpaCy's part-of-speech tagging tool to extract all nouns from the dataset's text. Subsequently, we create a set of nouns, denoted as S_n, comprising those nouns that occur more than 10 times. For a given image v_i, if the frequency of a particular noun in its k nearest-neighbor text surpasses a threshold p, we conclude that the image v_i contains the object represented by that noun, denoted as $obj(v_i)$:

$$obj(v_i) = \{n | \frac{\sum_{t_j \in V2T_k(v_i)} \mathbb{1}_{(n \in t_j)}}{k} > p, n \in S_n\}, \tag{7}$$

where the binary indicator $\mathbb{1}_{(n \in t_j)}$ equals 1 if the noun n appears in text t_j and 0 otherwise. The values of k and p may vary depending on the specific characteristics of the dataset, and we can confirm these values by conducting a rapid manual check to ensure the extraction of objects from the image.

We transform $obj(v_i)$ by inserting spaces between the nouns, resulting in the revised text t_i^r. During training, it is crucial to determine which noise texts require replacement. We set a threshold α. For image-text pairs (v_i, t_i) with a similarity of $(\mathbf{v}_i^g)^T \mathbf{t}_i^g \leq \alpha$, we replace the text t_i with the corresponding revised text t_i^r. Inspired by curriculum learning[4, 39, 40, 42], we progressively introduces more complex and challenging samples throughout the training process. Specifically, we replace all text during epoch 0, i.e., $\alpha = \max_{i \in [1,D]} \{(\mathbf{v}_i^g)^T \mathbf{t}_i^g\}$. After half of the total training epochs, we do not replace any text, i.e., $\alpha \leq \min_{i \in [1,D]} \{(\mathbf{v}_i^g)^T \mathbf{t}_i^g\}$. We employ a linear decay for α:

$$\alpha = \frac{e - E/2}{-E/2} \max_{i \in [1,D]} \{(\mathbf{v}_i^g)^T \mathbf{t}_i^g\} + \frac{e}{E/2} \min_{i \in [1,D]} \{(\mathbf{v}_i^g)^T \mathbf{t}_i^g\}, \tag{8}$$

where E is the total number of epochs and e is the number of the current epoch. This strategy ensures that the model is not disrupted by noise during the initial training, leading to a well-optimized starting point. Once alignment performance is sufficiently high, the model can gradually comprehend more complex original text.

3.3.2 GP-S for Uni-Modal. For uni-modal alignment, the focus is to sample image pairs with both structural and semantic similarity, we define semantic-equivalent images SE_i for v_i as follows:

$$SE_i = V2V_{k_1}(v_i) \cap T2V_{k_2}(t_i), \tag{9}$$

where k_1 and k_2 are parameters that control the range of neighbors. Here, information from $V2V$ requires similarity in various aspects such as composition of the images, while information from $T2V$ places greater emphasis on the semantic similarity of the images. When there are no positive images for v_i in the dataset, the intersection of the two sets would be empty, and we set SE_i as a set containing randomly augmented versions of v_i.

In each training epoch, we randomly select a positive image v_i^{pos} from SE_i and compute the uni-modal contrastive loss defined as follows:

$$\mathcal{L}_{\text{uni}}(v_i, v_i^{\text{pos}}) = -log \frac{\exp((\mathbf{v}_i^{\text{cls}})^T \mathbf{v}_i^{\text{pos,cls}} / \tau)}{\sum_{j=1}^{N} \exp((\mathbf{v}_i^{\text{cls}})^T \mathbf{v}_j^{\text{cls}} / \tau)}, \tag{10}$$

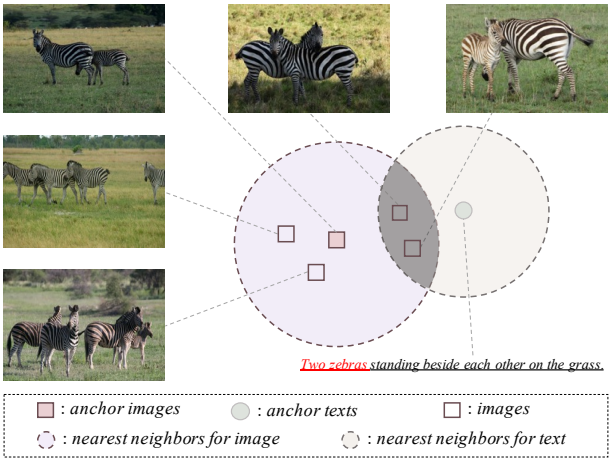

Figure 4: Illustration of GP-S for uni-modal. We identify positive images through the intersection between $T2V$ and $V2V$, allowing more precise sampling.

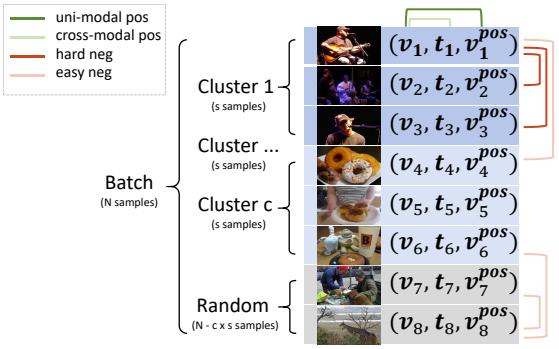

Figure 5: Illustration of GN-S. We present an example of the composition of a batch, where $N = 8$, $c = 2$, and $s = 3$. Samples within the same cluster serve as hard negative pairs for each other.

where τ is a learnable temperature parameter, N is the batch size.

For the input (v_i, t_i, v_i^{pos}), the final objective $\mathcal{L}$ is achieved by adding the uni-modal contrastive loss to the raw training loss of the model $\mathcal{L}_f$ outlined in its original paper:

$$\mathcal{L}(v_i, t_i, v_i^{\text{pos}}) = \mathcal{L}_f(v_i, t_i) + \mathcal{L}_{\text{uni}}(v_i, v_i^{\text{pos}}). \tag{11}$$

3.4 Global Negative Sampling (GN-S)

Previous works[2, 25, 33] have demonstrated the benefits of sampling hard negatives in VLP. However, in-batch mining is constrained by the batch size. In this section, we introduce global negative sampling, enabling us to acquire more challenging negative samples from the entire dataset.

An intuitive approach is to utilize k nearest neighbors as hard negatives. However, to avoid potential positives in the negative neighbors, the value of k needs to be set relatively high, which in turn results in an increase in computational costs. To enhance efficiency, we opt for the K-means algorithm[14] for clustering, where any pair of samples within the same cluster serve as a hard negative pair. Specifically, we utilize Faiss[17] to partition the dataset into K clusters, with each image-text pair (v_i, t_i) represented by its image feature $\mathbf{v}_i^g$.

As recent VLP models have incorporated in-batch hard negative mining, introducing our GN-S only requires placing global-level hard negatives into the same batch. Specifically, when a batch of size N is constructed, we randomly select c clusters from all K clusters and s samples from each cluster. The remaining $N - c \times s$ samples are randomly chosen from the entire dataset, as shown in Figure 5. By employing this approach, the s samples from the same cluster act as hard negatives for each other, resulting in a notable increased number of hard negatives compared to batches formed randomly. We observe that the model performs best when $c \times s \approx N/4$. This is done to maintain diversity within the batch, with still having $3N/4$ completely random samples, thus preventing the model from having difficulty distinguishing the 'easy' negatives.

3.5 Scalablity

Our framework is highly effective and efficient. Thanks to Faiss[17], computations such as 500-nearest-neighbor search and 1000-means clustering on a dataset containing 5 million image-text pairs can be completed within 2 hours, utilizing only 2 NVIDIA GeForce RTX-3090 GPUs. As a comparison, training the VLP model ALBEF[25] using 8 A100 GPUs takes 60 hours. **The additional computations introduced by our framework do not exceed 10% of the training time.** Regarding storage overhead, only the identifiers of the neighbors are stored, such as filenames. The storage space required is significantly smaller compared to that of the original image-text pairs.

When applying our framework to a large-scale dataset, we advise dividing the dataset randomly into groups of 5 million data pairs each and sampling neighbors only within each group. This approach has two advantages. First, our experiments have already demonstrated that the neighbor information computed on the 5 million dataset is already sufficient for our global positive and negative sampling, thereby obviating the need for any alteration in hyperparameters. Second, this processing method can be extended to datasets of any size, and the additional costs would also not exceed 10% of the training time.

4 Experiment

4.1 Implementation Details

Our GPN-S modifies the existing VLP model's sampling strategy to enhance their performance. We implement GPN-S on ALBEF [25] and TCL [44] model, referring to them as the backbone models. They share similar model architectures and pre-train datasets. Specifically, we utilized their base-size versions. The vision encoder in both models is a ViT-B/16 with 12 layers, and both the text encoder and the fusion encoder are implemented using a 6-layer transformer. The pre-training dataset includes COCO[27], Visual Genome[20], Conceptual Captions[35] and SBU Captions[31], comprising approximately 4 million images and 5.1 million image-text pairs.

Table 1: Results on fine-tune(FT) and zero-shot(ZS) retrieval tasks.

Method	MSCOCO-FT							MSCOCO-ZS						
	TR@1	TR@5	TR@10	IR@1	IR@5	IR@10	RSUM	TR@1	TR@5	TR@10	IR@1	IR@5	IR@10	RSUM
CLIP[33]	-	-	-	-	-	-	-	58.4	81.5	88.1	37.8	62.4	72.2	400.4
ViLT[19]	61.5	86.3	92.7	42.7	72.9	83.1	439.2	56.5	82.6	89.6	40.4	70.0	81.1	420.2
UNITER[10]	64.4	87.4	93.1	50.3	78.5	87.2	460.9	64.1	87.7	93.3	48.8	76.7	85.8	456.4
CoCa[46]	-	-	-	-	-	-	-	63.8	84.7	90.7	47.5	72.4	80.9	440.0
VLMo[2]	74.8	93.1	96.9	57.2	82.6	89.8	494.4	-	-	-	-	-	-	-
METER[12]	76.2	93.2	96.8	57.1	82.7	90.1	496.1	-	-	-	-	-	-	-
ALBEF[25]	73.1	91.4	96.0	56.8	81.5	89.2	488.0	68.7	89.5	94.7	50.1	76.4	84.5	463.9
+ours	76.0	92.9	96.7	58.5	82.7	89.6	496.4	71.8	91.1	95.2	53.8	79.3	87.1	478.3
	(+2.9)	(+1.5)	(+0.7)	(+1.7)	(+1.2)	(+0.4)	(+8.4)	(+3.1)	(+1.6)	(+0.5)	(+3.7)	(+2.9)	(+2.6)	(+14.4)
TCL[44]	75.6	92.8	96.7	59.0	83.2	89.9	497.2	71.4	90.8	95.4	53.5	79.0	87.1	477.2
+ours	77.3	93.7	96.8	59.9	83.6	90.2	501.5	72.5	91.8	95.7	55.3	80.5	88.0	483.8
	(+1.7)	(+0.9)	(+0.1)	(+0.9)	(+0.4)	(+0.3)	(+4.3)	(+1.1)	(+1.0)	(+0.3)	(+1.8)	(+1.5)	(+0.9)	(+6.6)

Table 2: Results on other vision-language tasks.

Method	VQA		$NLVR^2$		SNLI-VE	
	test-dev	test-std	dev	test-P	val	test
ViLT	71.26	-	75.70	76.13	-	-
UNITER	72.70	72.91	77.18	77.85	78.59	78.28
UNIMO	73.79	74.02	-	-	80.00	79.10
VLMo	76.64	76.89	82.77	83.34	-	-
METER	77.68	77.64	82.33	83.05	80.86	81.19
ALBEF	74.54	74.70	80.24	80.50	80.14	80.30
+ours	75.15	75.12	80.73	80.93	80.53	80.41
	(+0.61)	(+0.42)	(+0.49)	(+0.43)	(+0.39)	(+0.11)
TCL	74.90	74.92	80.54	81.33	80.51	80.29
+ours	75.46	75.58	81.14	81.44	80.63	80.51
	(+0.56)	(+0.66)	(+0.60)	(+0.11)	(+0.12)	(+0.22)

By default, we use the backbone model as the off-the-shelf pre-trained encoder. For instance, when applying our methods to ALBEF, we obtain the pre-trained checkpoint from their official GitHub repository for neighbor calculation. We empirically set the parameters as follows: $k = 10, p = 0.3$ for calculating $obj(v_i)$, $k_1 = 5, k_2 = 500$ for selecting positive image pairs, $c = 40, s = 3$ for constructing batches of size $N = 512$ in GN-S, with the number of clusters $K = 1000$. We maintain all settings and training details from the original backbone models, with one modification: we train TCL for 50 epochs, rather than 30, to achieve better convergence. All experiments are conducted on 8 NVIDIA A100 GPUs.

4.2 Downstream Tasks

To evaluate the effectiveness of our proposed method, we adapt the pre-trained model to various downstream vision-and-language (V+L) tasks, ensuring consistency in all settings with those of the backbone models[25, 44], as below:

Image-Text Retrieval. We evaluate both fine-tune and zero-shot retrieval performance on the MSCOCO[27] dataset and employ the widely used Karpathy split [18], which comprises 5000 images along with their corresponding 25010 texts. We consider two tasks: image-to-text retrieval (TR), which involves finding the corresponding text for a given image, and text-to-image retrieval (IR), which requires finding the corresponding image with a given text.

Visual Question Answering (VQA). VQA aims to answer natural language questions about images. We fine-tune our model on training and development set of VQAv2[1] and Visual Genome[20] VQA data, and evaluate on the test-dev and test-std set of VQAv2. We follow ALBEF and TCL, employing a decoder to generate answers from a pool of 3192 candidates.

Natural Language for Visual Reasoning for Real ($NLVR^2$) [36]. $NLVR^2$ presents the task of determining whether a natural language sentence is true about a pair of images. We follow ALBEF and TCL, extending the fusion encoder to enable reasoning over two images and adding a fully-connected layer to predict if the sentence is true or false.

Visual Entailment (SNLI-VE) [43]. SNLI-VE is a novel inference task based on the Stanford Natural Language Inference corpus and Flickr30k dataset. Given a real world image premise and a natural language hypothesis, the goal is to determine if the hypothesis can be concluded given the information provided by image. We follow ALBEF and TCL, considering it as a three-classes classification problem because the relationship could be entailment, neutral or contradiction.

For image-text retrieval, the R@n (Recall at n) measures the proportion of correct answers included within the top n retrieved results. The RSUM metric is the sum of TR@1, TR@5, TR@10, IR@1, IR@5, and IR@10. For VQA, $NLVR^2$ and SNLI-VE, we report the average accuracy on the test dataset.

4.3 Main Results

This subsection compares the performance of models with and without the GPN-S framework to demonstrate its effectiveness.

Table 3: Performance impact of individual sampling strategies on downstream tasks. For retrieval task, we report the RSUM metric.

Module	MSCOCO ZS	MSCOCO FT	VQA test-dev	NLVR^2 test-P	SNLI-VE test
ALBEF	463.9	488.0	74.54	80.50	80.30
+GP-S(Uni)	474.0	492.7	74.74	80.74	80.25
+GP-S(Cross)	474.0	492.4	74.81	80.87	80.23
+GP-S(Uni+Cross)	475.1	492.8	75.07	80.91	80.36
+GN-S	475.9	493.0	75.05	80.90	80.28
+GPN-S(FULL)	**478.3**	**496.4**	**75.15**	**80.93**	**80.41**
TCL	477.2	497.2	74.90	81.33	80.29
+GP-S(Uni)	481.8	499.2	75.14	81.33	80.45
+GP-S(Cross)	481.7	499.3	75.21	81.35	80.52
+GP-S(Uni+Cross)	482.2	499.2	75.43	81.38	**80.55**
+GN-S	483.1	501.2	75.32	81.37	80.51
+GPN-S(FULL)	**483.8**	**501.5**	**75.46**	**81.44**	80.51

4.3.1 Image-Text Retrieval. We evaluate the performance of the model in both fine-tune and zero-shot settings, with the results shown in Table 1.

In both settings, the model incorporating GPN-S significantly outperforms the original backbone model across all metrics. Notably, the improvement is more pronounced in the zero-shot setting, with increases of +14.4(3.1%) for ALBEF and +6.6(1.4%) for TCL in terms of RSUM metrics. This enhancement indicates that GPN-S more effectively aligns images and text, and the more substantial increase in R@1 compared to R@5 and R@10 suggests that GPN-S enhances the model's ability to discern fine-grained differences.

4.3.2 VQA, NLVR^2 and SNLI-VE. We conduct fine-tuning on other downstream tasks requiring deep image-text interaction understanding. Table 2 shows the experimental results. Our method consistently outperforms the corresponding backbone models.

Specifically, on the VQA task, we achieve an average improvement of +0.52(+0.7%) and +0.61(+0.8%) for the ALBEF and TCL backbones, respectively. For the NLVR^2 task, our method outperforms the backbone by an average of 0.41(+0.5%). For the SNLI-VE task, our method yields an average improvement of 0.21(+0.3%) over the backbone. These gains are attributed solely to the improved pre-training alignment, as the fine-tuning process remains unchanged.

4.4 Ablation Study

4.4.1 Sampling Strategies. The effectiveness of each sampling strategy is validated through experiments, summarized in Table 3. Each strategy demonstrates improvement across all tasks, except SNLI-VE, compared to the backbone model. Combining two global positive sampling strategies yields better results than employing a single strategy, indicating that uni-modal and cross-modal positive samplings are complementary. Implementing all three strategies leads to the highest performance in most metrics, confirming their collective benefit.

4.4.2 Off-The-Shelf Encoder. We investigate the effectiveness of various off-the-shelf models. The results on the ALBEF backbone are detailed in Table 4. By default, we use the backbone model that is not trained by our framework as the off-the-shelf model (so 'ALBEF' in the table denotes the default method). For comparison, we also utilize CLIP ViT-B/32 and TCL models as alternative off-the-shelf encoders, sourced from their official repositories. Our findings indicate that higher-performing off-the-shelf models yield better results in zero-shot retrieval tasks, while maintaining similar levels of performance in other downstream tasks. Given the significant improvements observed across all models, it is advisable to choose the best available off-the-shelf encoder for applying the GPN-S framework.

Table 4: Performance impact of different off-the-shelf encoders. For retrieval task, we report the RSUM metric.

Encoder	MSCOCO ZS	MSCOCO FT	VQAv2 test-dev	NLVR^2 test-P	SNLI-VE test
-	463.9	488.0	74.54	80.50	80.30
CLIP	477.1	**496.4**	**75.27**	81.11	80.23
ALBEF	478.3	**496.4**	75.15	80.93	**80.41**
TCL	**479.1**	496.3	75.12	**81.20**	80.22

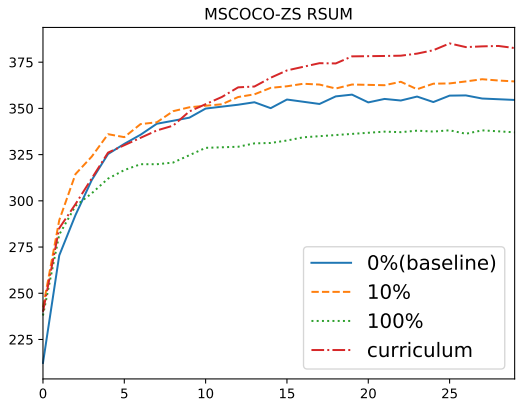

Figure 6: The RSUM metric for zero-shot retrieval on the MSCOCO dataset at each epoch during the training process (x-axis represents the number of training epochs). The 0% line represents the original ALBEF model. The 10% line shows a fixed α scenarios where text is replaced if its similarity with the image falls within the lowest 10% of the dataset. The 100% line represents that all text is replaced. The curriculum line illustrates our curriculum-based scheduling for α.

4.5 GP-S for Cross-modal on Noisier Dataset

The GP-S method was proposed to address the impact of textual noise. However, in previous sections, we used relatively clean academic datasets for a fair comparison. In this section, we replace the pre-training dataset with a noisier random 3M subset of CC12M and use GP-S for cross-modal techniques without modifying the parameters k and p. We report the RSUM metric for zero-shot retrieval

Figure 7: Two examples are shown for visualization. The first column displays the original image-text pairs, while the following columns present the positive and negative samples obtained via our sampling strategy.

on the MSCOCO dataset, as shown in Figure 6. We observe that, on this dataset, we achieve a better improvement (RSUM improvement from 354.6 to 382.6, a 7.9% increase). In contrast to a fixed α, curriculum learning for scheduling α proves to be the most effective approach. We believe that curriculum training, progressing from easy to hard, prevents the model from getting stuck in local optima.

As pre-training datasets in practical applications continue to grow, they will inevitably contain more noise. As the first work to introduce positive sampling in VLP, we hope this experiment demonstrates the importance of positive sampling in VLP.

4.6 Visualization of Positives and Negatives

For a clearer understanding of our sampling strategy, we provide visualizations of the positive and negative samples identified by GPN-S, as illustrated in Figure 7.

In the first example (first row), the text mentions *"saturday afternoon"* which is not visually represented in the image. We obtain information through neighbors and revise the text accordingly. Moreover, We found neighbor images with similar semantic information and encouraged them to be close to each other in the embedding space. To distinguish hard negative samples within the same cluster, the model is required to fully comprehend the sentence *"takes control of the ball from people,"* rather than solely recognizing *"people"* and *"ball"*. In the second example, cross-modal positive sampling successfully captures the key information *"snow"* which is not explicitly mentioned in the original text. All positive images contain the information *"skating over a red platform"* while negative samples lack certain shared features with the anchor, yet are sufficiently challenging to enable the model to achieve fine-grained discrimination.

5 Conclusion

This paper introduces the curriculum-based Global Positive-Negative Sampling (GPN-S) framework, a novel approach that enhances the sampling strategy of Vision-Language Pre-training (VLP) models by including both positive and negative pairs at a global level. The GPN-S framework leverages curriculum learning to progressively introduce more complex and informative samples during training. This process identifies globally semantically similar neighbors as positive samples, thereby improving the quality of representations in both cross-modal and uni-modal scenarios. Simultaneously, it encourages the model to differentiate more challenging negative samples within the same cluster. Experimental results validate the effectiveness of the GPN-S framework, demonstrating significant improvements across various VLP models on widely used benchmarks, and highlighting the potential of sampling methods in advancing the state-of-the-art in vision-language pre-training.

Acknowledgments

This work was supported by the National Key Research and Development Program of China No. 2023YFF1205001, National Natural Science Foundation of China (No. 62250008, 62222209, 62102222), Beijing National Research Center for Information Science and Technology under Grant No. BNR2023RC01003, BNR2023TD03006, and Beijing Key Lab of Networked Multimedia.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.

[2] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, and Furu Wei. 2021. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358* (2021).

[3] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* 35, 8 (2013), 1798–1828.

[4] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*. 41–48.

[5] Jaeseok Byun, Taebaek Hwang, Jianlong Fu, and Taesup Moon. 2022. Grit-vlp: Grouped mini-batch sampling for efficient vision and language pre-training. In *European Conference on Computer Vision*. Springer, 395–412.

[6] Hong Chen, Yudong Chen, Xin Wang, Ruobing Xie, Rui Wang, Feng Xia, and Wenwu Zhu. 2021. Curriculum disentangled recommendation with noisy multi-feedback. *Advances in Neural Information Processing Systems* 34 (2021), 26924–26936.

[7] Houlun Chen, Xin Wang, Xiaohan Lan, Hong Chen, Xuguang Duan, Jia Jia, and Wenwu Zhu. 2023. Curriculum-listener: Consistency-and complementarity-aware audio-enhanced temporal sentence grounding. In *Proceedings of the 31st ACM International Conference on Multimedia*. 3117–3128.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[9] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. 2020. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297* (2020).

[10] 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: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*. Springer, 104–120.

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

[12] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. 2022. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18166–18176.

[13] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. 2021. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9588–9597.

[14] John A Hartigan and Manchek A Wong. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)* 28, 1 (1979), 100–108.

[15] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. 2024. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14271–14280.

[16] 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 *International Conference on Machine Learning*. PMLR, 4904–4916.

[17] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* 7, 3 (2019), 535–547.

[18] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3128–3137.

[19] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*. PMLR, 5583–5594.

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

[21] Xiaohan Lan, Yitian Yuan, Hong Chen, Xin Wang, Zequn Jie, Lin Ma, Zhi Wang, and Wenwu Zhu. 2023. Curriculum multi-negative augmentation for debiased video grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 1213–1221.

[22] Haoyang Li, Xin Wang, and Wenwu Zhu. 2023. Curriculum graph machine learning: A survey. *arXiv preprint arXiv:2302.02926* (2023).

[23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597* (2023).

[24] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*. PMLR, 12888–12900.

[25] 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. *Advances in neural information processing systems* 34 (2021), 9694–9705.

[26] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*. Springer, 121–137.

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

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems* 36 (2024).

[29] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems* 32 (2019).

[30] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2630–2640.

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

[32] Yijian Qin, Xin Wang, Ziwei Zhang, Hong Chen, and Wenwu Zhu. 2024. Multi-task graph neural architecture search with task-aware collaboration and curriculum. *Advances in neural information processing systems* 36 (2024).

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[34] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114* (2021).

[35] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. 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)*. 2556–2565.

[36] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2018. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491* (2018).

[37] Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490* (2019).

[38] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. 2020. What makes for good views for contrastive learning? *Advances in neural information processing systems* 33 (2020), 6827–6839.

[39] Xin Wang, Yudong Chen, and Wenwu Zhu. 2021. A survey on curriculum learning. *IEEE transactions on pattern analysis and machine intelligence* 44, 9 (2021), 4555–4576.

[40] Xin Wang, Zirui Pan, Yuwei Zhou, Hong Chen, Chendi Ge, and Wenwu Zhu. 2023. Curriculum co-disentangled representation learning across multiple environments for social recommendation. In *International Conference on Machine Learning*. PMLR, 36174–36192.

[41] Xin Wang, Yuwei Zhou, Hong Chen, and Wenwu Zhu. 2024. Curriculum Learning: Theories, Approaches, Applications, Tools, and Future Directions in the Era of Large Language Models. In *Companion Proceedings of the ACM on Web Conference 2024*. 1306–1310.

[42] Zihao Wu, Xin Wang, Hong Chen, Kaidong Li, Yi Han, Lifeng Sun, and Wenwu Zhu. 2023. Diff4rec: Sequential recommendation with curriculum-scheduled diffusion augmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9329–9335.

[43] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706* (2019).

[44] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. 2022. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15671–15680.

[45] Yang Yao, Xin Wang, Yijian Qin, Ziwei Zhang, Wenwu Zhu, and Hong Mei. 2024. Data-augmented curriculum graph neural architecture search under distribution shifts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38.

16433–16441.

[46] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917* (2022).

[47] Yuwei Zhou, Hong Chen, Zirui Pan, Chuanhao Yan, Fanqi Lin, Xin Wang, and Wenwu Zhu. 2022. Curml: A curriculum machine learning library. In *Proceedings of the 30th ACM International Conference on Multimedia*. 7359–7363.

[48] Yuwei Zhou, Zirui Pan, Xin Wang, Hong Chen, Haoyang Li, Yanwen Huang, Zhixiao Xiong, Fangzhou Xiong, Peiyang Xu, Wenwu Zhu, et al. [n. d.]. CurBench: Curriculum Learning Benchmark. In *Forty-first International Conference on Machine Learning*.

[49] Yuwei Zhou, Xin Wang, Hong Chen, Xuguang Duan, Chaoyu Guan, and Wenwu Zhu. 2022. Curriculum-nas: Curriculum weight-sharing neural architecture search. In *Proceedings of the 30th ACM International Conference on Multimedia*. 6792–6801.

[50] Yuwei Zhou, Xin Wang, Hong Chen, Xuguang Duan, and Wenwu Zhu. 2023. Intra-and Inter-Modal Curriculum for Multimodal Learning. In *Proceedings of the 31st ACM International Conference on Multimedia*. 3724–3735.