Negative Pre-aware for Noisy Cross-Modal Matching

Xu Zhang[1†], **Hao Li**[1†], **Mang Ye**[2*]

[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China
[2]School of Computer Science, Wuhan University
{xuzhang.xoe, 18th.leolee, mangye16}@gmail.com

Abstract

Cross-modal noise-robust learning is a challenging task since noisy correspondence is hard to recognize and rectify. Due to the cumulative and unavoidable negative impact of unresolved noise, existing methods cannot maintain a stable performance when the noise increases. In this paper, we present a novel Negative Pre-aware Cross-modal (NPC) matching solution for large visual-language model fine-tuning on noisy downstream tasks. It is featured in two aspects: (1) For noise recognition and resistance, previous methods usually directly filter out a noise subset, we propose to estimate the negative impact of each sample. It does not need additional correction mechanisms that may predict unreliable correction results, leading to self-reinforcing error. We assign a confidence weight to each sample according to its negative impact in the training process. This adaptively adjusts the contribution of each sample to avoid noisy accumulation. (2) For maintaining stable performance with increasing noise, we utilize the memorization effect of DNNs by maintaining a memory bank. Specifically, we apply GMM to select high-confident clean samples as the memory entry, where the memory entry is used to estimate the negative impact of each sample. Since clean samples are easier distinguished by GMM with increasing noise, the memory bank can still maintain high quality at a high noise ratio. Compared to the correction mechanism focusing on noise samples, memory bank-based estimation is more robust, which makes the model performance stable on noisy datasets. Extensive experiments demonstrate that our method significantly improves matching accuracy and performance stability at increasing noise ratio. Our approach also surpasses the state-of-the-art methods by a large margin. The code is available at: https://github.com/ZhangXu0963/NPC.

Introduction

Cross-modal matching aims to align different modalities (*e.g.*, text and image) within a common space and pair them based on similarity score. With the explosion of multimedia data, cross-modal matching has gained traction in both industry and academia, *e.g.*, text-to-image generation (Zhou et al. 2022; Ding et al. 2021), image captioning (Li et al. 2019b; Stefanini et al. 2022; Wang et al. 2023), and visual question answering (Lin et al. 2022; Lei et al. 2023).

*Corresponding author.
†These authors contributed equally to this work.

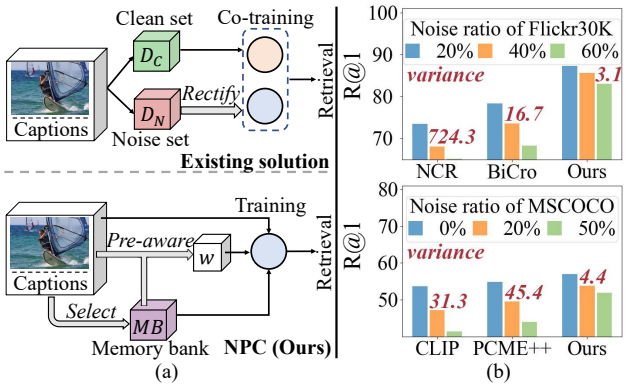


Figure 1: (a) Existing solution vs NPC. (b) The variance of R@1 of noise-robust learning and CLIP-based methods. A lower variance indicates that the method is more robust in the face of increasing noise.

These works have achieved promising performance by training on large-scale datasets. However, it is expensive to obtain a well-annotated dataset in practical scenarios. The manual-annotated datasets, *e.g.*, MSCOCO (Lin et al. 2014), Flickr30K (Young et al. 2014), and Conceptual Captions (Sharma et al. 2018), incorporate a significant number of inaccurate descriptions, namely noisy correspondence. Unlike noisy label in classification tasks, the noise here is mismatched cross-modal pairs which is more difficult to deal with, since involves both visual and textual modeling. Therefore, a series of approaches (Huang et al. 2021; Yang et al. 2023; Han et al. 2023) following the noise-rectify paradigm have been developed to counter the negative impact of the noise. These methods typically filter out the noise subset from the original training set, and address the noise issue through label correction. Nevertheless, the inherent flaw of the noise-rectify paradigm cannot maintain the performance stability in the existence of severe noise. As shown in Fig. 1(b), we compare the performance of different methods using R@1 metric, including noise-rectify based approaches (Huang et al. 2021; Yang et al. 2023), the CLIP-based approaches (Radford et al. 2021; Chun 2023) and our approach. We employ variance (var) of R@1 at different noise ratio to illustrate the "performance stabil-

ity". Obviously, noise-rectify based methods exhibit unstable performance with a considerably larger variance than ours. Additionally, CLIP-based methods also lack consistent performance with increasing noise, even though CLIP is a powerful pre-trained model. Most existing noise-rectify paradigms rely on collaborative rectifying with multiple models. Since the limitation of the rectifying mechanism, the matching performance under high-noise is unstable. In these works (Huang et al. 2021; Yang et al. 2023), the new labels are entirely estimated by DNN models. With high noise ratio, some indistinguishable noise correspondences are prone to be directly learned and remembered by DNNs, ultimately leading to a dramatic drop in performance under high-noise. Existing methods emphasize the "discrimination learning" ability but ignore the "stability". In our opinion, two essential abilities are required for noise-robust learning for large visual-language model fine-tuning on noisy downstream tasks: 1) *the ability to distinguish between noisy and clean samples*, 2) *maintain the stability of "discriminative learning" with increasing noise*.

To address aforementioned challenges, we propose a novel approach named Negative Pre-aware Cross-modal (NPC) matching. NPC adopts a unique Negative Pre-aware (NP) paradigm for robust learning. Unlike previous paradigms that mainly focus on noise filtering or correction, the NP paradigm adaptively assesses the potential negative impact of each sample before the model learning (see Fig. 1(a)). DNNs tends to prioritize learning easy samples over noisy and challenging ones (Arpit et al. 2017; Xia et al. 2021). With gradually fitting noise samples, the model begins to generate incorrect predictions (Liu et al. 2020). In other words, once the model learned a noise pair, fitting certain specific clean samples becomes more challenging. These clean samples usually have images or texts that are similar to the noise pair.

Inspired by this phenomenon, our NPC uses easy-distinguishable clean samples to estimate negative impacts. We rigorously choose a reliable clean subset from the training data by using Gaussian Mixture Model (Li, Socher, and Hoi 2020; Permuter, Francos, and Jermyn 2006) to fit the loss distribution of each pair. And high-confident clean samples are maintained in a Memory Bank (MB), which is used to assist the model in estimating negative impact before to fully model training. A small confidence weight will be assigned to high-negative samples.

The main contributions are summarized as follows:

- We highlight the challenge with large visual-language model fine-tuning on noisy downstream tasks, *i.e.*, how to achieve robust learning in cross-modal matching with the increasing amount of noise.
- We introduce the Negative Pre-aware Cross-modal (NPC) matching paradigm by establishing a memory bank for negative impact estimation. We employ the assistance of memory entries to allocate confidence weights (w) to the samples. These components constitute the cornerstones to achieving stable and highly noise-resistant performance.
- Extensive experiments are conducted on two manual-annotated datasets and a real-world dataset, showcasing the NPC's superiority over the state-of-the-art methods. Moreover, with the increasing noise, both quantitative and qualitative results affirm that NPC demonstrates notably higher performance stability compared to previous methods.

Related Works

Image-text Matching

Typical image-text matching methods align data from different modalities to measure similarity. Early works (Faghri et al. 2018; Song and Soleymani 2019; Wang et al. 2018; Qian et al. 2021) mainly focus on global alignments. Some prior works (Lee et al. 2018; Li et al. 2019a; Diao et al. 2021; Zhang et al. 2023) adopt attention mechanisms to achieve fine-grained local alignments. Subsequently, many works (Chun et al. 2021; Chun 2023; Li et al. 2022) devote to modeling the many-to-many relationships in image-caption pairs.

Recently, with the success of transformer-based vision and language models (Dosovitskiy et al. 2021; Devlin et al. 2019), vision-language pre-training (VLP) models, such as CLIP (Radford et al. 2021), have shown strong performance in multiple cross-modal tasks (Jiang and Ye 2023; You et al. 2023). Although VLP possesses impressive zero-shot ability, it still reveals vulnerabilities in training with noisy datasets on specific downstream tasks. In this paper, we employ CLIP as our backbone and introduce an anti-noise learning strategy.

Cross-modal Noise-robust Learning

Huang et al. (Huang et al. 2021) first tackle noise correspondences, which consider mismatched cross-modal pairs instead of incorrect annotations. Since then, several approaches (Han et al. 2023; Yang et al. 2023; Ye et al. 2022; Ye and Yuen 2020) have developed the noise-rectify process in various cross-modal tasks. They can be categorized into two groups: noise correction and noise re-weighting. Noise correction methods achieve robust learning by correcting the similarity (Han et al. 2023) or correspondence label (Huang et al. 2021; Yang et al. 2023) of noise pairs. The noise re-weighting methods (Qin et al. 2022) degrade the contribution of noise samples to achieve robust learning. All these methods require splitting a noise subset from the original training dataset. Subsequently, they proceed with rectification within this subset. Nonetheless, as noise increases, the imprecise subset division and inaccurate rectification can amplify adverse effects. Different from these works, we sidestep the problem by forecasting per-sample negative impact following the novel NP paradigm.

Proposed Method

Preliminary

Problem Definition. Given a dataset $D = \{(I_i, T_i)\}_{i=1}^{N}$, where (I_i, T_i) is the i^{th} image-text pair, and N denotes the data size. The goal of image-text matching is to align the visual and textual modalities within a shared space to calculate

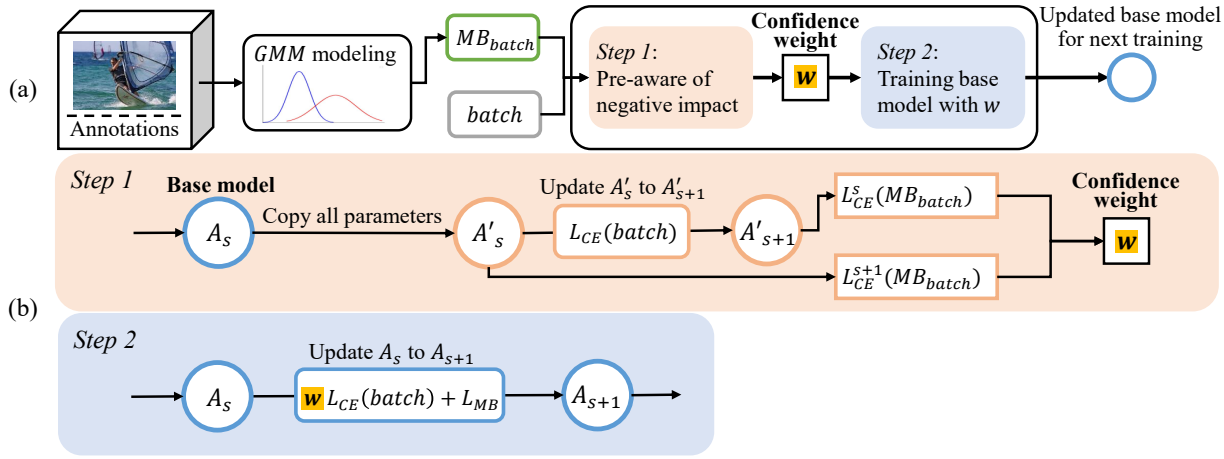


Figure 2: (a) Illustrating the NPC training pipeline. Given a batch of image-text pairs, we select their corresponding memory entries from a strict clean set divided by GMM as inputs. Then we optimize the base model in two steps: the first step aims to estimate the negative impact and obtain per-sample confidence weight w. The second step is training the base model with w. (b) Illustrating two training steps. We first share all parameters of the base model A_s to its siamese model A'_s. Then we train the model A'_s on the batch samples, obtaining the model A'_{s+1}. The negative impact of each sample can be calculated by comparing its loss of corresponding memory entry on A'_s and A'_{s+1}. If the loss on A'_{s+1} is higher than it on A'_s, this means the sample brings a negative impact to the model, and we will give it a low confidence weight. After the negative-aware process, the model A_s will be trained with the re-weight samples and memory bank, generating the robust target model A_{s+1}.

the similarity following Eq. 1,

$$S(I_i, T_j) = \frac{f(I_i) \cdot g(T_j)}{\|f(I_i)\| \cdot \|g(T_j)\|}, \tag{1}$$

where $f(\cdot)$ and $g(\cdot)$ serve as feature extractors for two modalities. Generally, positive pairs exhibit higher similarity scores, whereas negative pairs show lower similarity scores.

Revisiting CLIP-based Solution. With the emergence of the VLP model CLIP (Radford et al. 2021) as a compelling option for cross-modal downstream tasks, we employ CLIP as the pre-trained backbone for the proposed NPC approach. CLIP enhances visual and textual feature extractors through the minimization of the symmetric cross-entropy loss $\mathcal{L}_{CE}(I_i, T_i)$, which is defined as follows:

$$\begin{aligned} \mathcal{L}_{CE}(I_i, T_i) &= CE(I_i, T_i) + CE(T_i, I_i), \\ CE(x_i, y_i) &= -\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{\exp(S(x_i, y_i))}{\sum_{j=1}^{N}\exp(S(x_i, y_j))}\right). \end{aligned} \tag{2}$$

However, Eq. 2 works effectively based on the assumption that (I_i, T_i) constitutes a positive pair. Yet, when (I_i, T_i) is a noise correspondence, relying solely on Eq. 2 can lead to a substantial detrimental impact on the model.

Fig. 1 provides a clear visual representation, demonstrating that when the noise ratio rises from 20% to 60%, the CLIP's R@1 performance experiences a steep decline from 82.3% to 66.3%. Therefore, the NPC approach is introduced to enhance the stability and robustness of pre-trained models in tackling noise challenges. The training pipeline, depicted in Fig. 2, comprises the two main components that will be elaborated upon in the subsequent section.

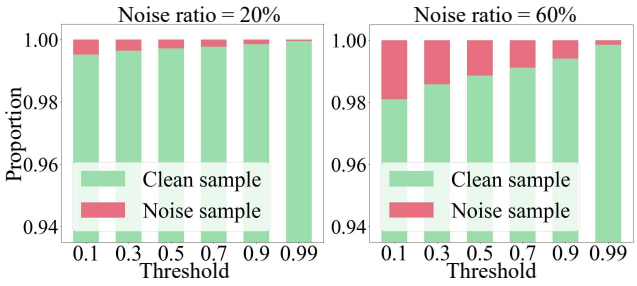


Figure 3: The proportion of noise and clean samples in the clean set, obtained through GMM at different thresholds τ. Generally, samples with the posterior probability of $p_i \geq \tau$ are included in the clean set. Inevitably, there are some noise samples in it. The threshold $\tau = 0.99$ ensures that the clean set selected from either the low (e.g. 20%) or the high (e.g. 60%) noise ratio training set is virtually noise-free.

Memory Bank Construction

We propose to estimate the negative impact of each sample brought to the model during the training process. A direct approach is to evaluate the performance changes of the model before and after training. Limited by the high cost of evaluating on the test set, we construct corresponding evaluation entries for each sample, which together form a Memory Bank (MB). Concretely, we select these entries from a reliable clean set to guarantee the accuracy of evaluation. Since DNN tends to learn the easy patterns before noisy and hard patterns (Arpit et al. 2017; Xia et al. 2021), clean samples typically exhibit lower loss values than the noisy or hard

ones. Based on this, we leverage the difference in loss distribution among samples to discern clean pairs. Following NCR (Huang et al. 2021), we utilize a two-component Gaussian Mixture Model to fit the distribution of per-sample loss in the training dataset:

$$p(z|\theta) = \sum_{k=1}^{K} \alpha_k \phi(z|\theta_k), \tag{3}$$

where α_k represents the mixture coefficient, and $\phi(z|\theta_k)$ denotes the probability density of the k^{th} component. The posterior probability computed by Eq. 4 serves as the clean probability p_i for the i^{th} sample.

$$p_i = p(\theta_k|z_i) = \frac{p(\theta_k)p(z_i|\theta_k)}{p(z_i)}. \tag{4}$$

Here, θ_k refers to the Gaussian component with a lower mean. The samples with $p_i \geq \tau$ are considered clean, as indicated in Eq. 5.

$$D_c = \{(I_j, T_j)|p_j \geq \tau\}. \tag{5}$$

Fig. 3 illustrates the proportion of noise and clean samples in the selected dataset D_c with varied threshold τ. We perform the strict selection using $\tau = 0.99$ to obtain the clean set D_c, practically devoid of noise. Strict selection is a prerequisite to ensure the reliability of the memory bank.

Then, we need to select evaluation entries in the strict clean set for each sample to construct the memory bank. For each pair (I_i, T_i) in the training set, we first select an image-text pair (I_i^I, T_i^I) from D_c for I_i, where the image in this pair (I_i^I) exhibits the highest cosine similarity (Eq. 2) with I_i. Similarly, we also choose an image-text pair (I_i^T, T_i^T) for T_i. The constructed memory bank can be defined as $MB = \{(I_i^I, T_i^I), (I_i^T, T_i^T)\}_{i=1}^{N}$.

Pre-aware of the Negative Impact

An intuitive fact is that when the model learns a noisy sample, its prediction accuracy of related clean samples will be declined. Therefore, after a sample is trained, we can determine its negative impact degree through the model performance on related clean samples. To estimate the negative impact of each sample, we have built the related clean evaluation entries for each sample, which together form a Memory Bank (MB).

During the batch with the size of m training shown in Fig. 2, both batch data and their corresponding memory entry set $MB_{batch} = \{b_1, b_2, \ldots, b_m\}$ are inputted into the model simultaneously. In the initial phase of each batch training, the base model A shares all parameters with A'. It's worth noting that the models A and A' update separately and independently. The purpose of A' is to perceive the negative impact of each sample in the batch by assessing the performance changes of the model on MB_{batch} after training. We utilize the loss to denote the performance, *i.e.*, the low loss almost means the model performs well on MB_{batch}. For the image-text pair (I_k, T_k), the losses of its evaluation entry b_k on both i2t and t2i can be computed by:

$$\begin{aligned} p_k &= CE(I_k^I, T_k^I) + CE(I_k^T, T_k^T), \\ q_k &= CE(T_k^I, I_k^I) + CE(T_k^T, I_k^T). \end{aligned} \tag{6}$$

Denote the model before and after training as A'_s and A'_{s+1}, respectively. The performance change of the model after the sample (I_k, T_k) trained can be calculated by:

$$r_k = \frac{1}{2}\left(\frac{p_k^s}{p_k^{s+1}} + \frac{q_k^s}{q_k^{s+1}}\right). \tag{7}$$

When $r_k < 1$, *i.e.*, the loss p_k and q_k increase after training. It means that the model's ability on predicting the correspondence of the clean pair related to the sample (I_k, T_k) is declined after training it. Thus, (I_k, T_k) has a negative impact on the model A'. We utilize the confidence weight w_k to quantify the negative impact of the pair (I_k, T_k) following Eq. 8. The sample with high negative impact (*i.e.*, low r_k) should correspond to a small confidence weight w_k.

$$w_k = \begin{cases} \tanh(r_k) & , r_k < 1 \\ 1 & , otherwise \end{cases} \tag{8}$$

The sample with $r_k < 1$ will bring a negative impact to the model. Thus, we will assign the confidence weight $w_k < 1$ computed by a tangent function for it. Similarly, for the samples with $r_k \geq 1$, we will assign the confidence weight $w_k = 1$. So far, in the batch, we can estimate the negative impact of each sample on the base model A.

Re-training

After negative impact evaluation, we need to re-train the model A to get the robust target model A_{s+1}. To avoid the detriment of the samples with a negative impact on the base model A, we re-weight the symmetric cross-entropy loss:

$$\mathcal{L}_{RCE} = \frac{1}{m}\sum_{k=1}^{m} w_k \mathcal{L}_{CE}(I_k, T_k). \tag{9}$$

For these detrimental samples, the labels are not reliable. To further mitigate the detriment of these unreliable labels to the model, we employ the related memory entries to help the model learn the correct correspondences (Eq. 10).

$$\mathcal{L}_{MB} = \frac{1}{m}\sum_{k=1}^{m}\left[\mathcal{L}_{CE}(I_k^I, T_k^I) + \mathcal{L}_{CE}(I_k^T, T_k^T)\right]. \tag{10}$$

Thus, the total objective function in the re-training process can be denoted as:

$$\mathcal{L}_{total} = \mathcal{L}_{RCE} + \mathcal{L}_{MB}. \tag{11}$$

Experiments

Experimental Setting

Datasets and Evaluation Metrics. The proposed NPC is evaluated on three benchmark datasets, MSCOCO (Lin et al. 2014), Flickr30K (Young et al. 2014), and CC120K:

- MSCOCO contains 123,287 images with 5 annotated captions per image. Following previous works (Huang et al. 2021), we use 113,287 images for training, 5,000 images for validation, and 5,000 images for testing.
- Flickr30K contains 31,783 images with 5 annotated texts per image. Following previous works (Huang et al. 2021), we use 29,783 images for training, 1,000 images for validation, and 1,000 images for testing.

noise	method	MSCOCO 1K						Flickr30K					
		image-to-text			text-to-image			image-to-text			text-to-image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
0%	SCAN	69.2	93.6	97.6	56.0	86.5	93.5	67.4	90.3	95.8	48.6	77.7	85.2
	SAF	76.1	95.4	98.3	61.8	89.4	95.3	73.7	93.3	96.3	56.1	81.5	88.0
	NCR	78.7	95.8	98.5	63.3	90.4	95.8	77.3	94.0	97.5	59.6	84.4	89.9
	DECL	79.1	96.3	98.7	63.3	90.1	95.6	79.8	94.9	97.4	59.5	83.9	89.5
	BiCro	79.1	96.4	98.6	63.8	90.4	96.0	81.7	95.3	98.4	61.6	85.6	90.8
	CLIP	79.9	95.1	98.1	65.0	90.3	98.1	86.2	97.6	99.2	72.9	92.3	96.0
	NPC	**82.2**	**96.5**	**98.7**	**68.3**	**92.0**	**98.7**	**87.9**	**98.1**	**99.4**	**75.0**	**93.7**	**97.2**
20%	SCAN	62.2	90.0	96.1	46.2	80.8	89.2	58.5	81.0	90.8	35.5	65.0	75.2
	SAF	71.5	94.0	97.5	57.8	86.4	91.9	62.8	88.7	93.9	49.7	73.6	78.0
	NCR	77.7	95.5	98.2	62.5	89.3	95.3	73.5	93.2	96.6	56.9	82.4	88.5
	DECL	77.5	95.9	98.4	61.7	89.3	95.4	77.5	93.8	97.0	56.1	81.8	88.5
	BiCro	78.8	**96.1**	**98.6**	63.7	90.3	95.7	78.1	94.4	97.5	60.4	84.4	89.9
	CLIP	75.0	93.1	97.2	58.7	86.1	97.2	82.3	95.5	98.3	66.0	88.5	93.5
	NPC	**79.9**	95.9	98.4	**66.3**	**90.8**	**98.4**	**87.3**	**97.5**	**98.8**	**72.9**	**92.1**	**95.8**
40%	SCAN	42.9	74.6	85.1	24.2	52.6	63.8	26.0	57.4	71.8	17.8	40.5	51.4
	SAF	13.5	43.8	48.2	16.0	39.0	50.8	7.4	19.6	26.7	4.4	12.2	17.0
	NCR	74.7	94.6	98.0	59.6	88.1	94.7	68.1	89.6	94.8	51.4	78.4	84.8
	DECL	75.6	95.5	**98.3**	59.5	88.3	94.8	72.7	92.3	95.4	53.4	79.4	86.4
	BiCro	77.0	**95.9**	**98.3**	61.8	89.2	94.9	74.6	92.7	96.2	55.5	81.1	87.4
	CLIP	70.7	91.7	96.2	54.7	83.4	96.2	76.2	93.3	96.5	59.4	85.0	90.9
	NPC	**79.4**	95.1	**98.3**	**65.0**	**90.1**	**98.3**	**85.6**	**97.5**	**98.4**	**71.3**	**91.3**	**95.3**
60%	SCAN	29.9	60.9	74.8	0.9	2.4	4.1	13.6	36.5	50.3	4.8	13.6	19.8
	SAF	0.1	0.5	0.7	0.8	3.5	6.3	0.1	1.5	2.8	0.4	1.2	2.3
	NCR	0.1	0.3	0.4	0.1	0.5	1.0	13.9	37.7	50.5	11.0	30.1	41.4
	DECL	73.0	94.2	**97.9**	57.0	86.6	93.8	65.2	88.4	94.0	46.8	74.0	82.2
	BiCro	73.9	**94.4**	97.8	58.3	87.2	93.9	67.6	90.8	94.4	51.2	77.6	84.7
	CLIP	67.0	88.8	95.0	49.7	79.6	95.0	66.3	87.3	93.0	52.1	78.8	87.4
	NPC	**78.2**	**94.4**	97.7	**63.1**	**89.0**	**97.7**	**83.0**	**95.9**	**98.6**	**68.1**	**89.6**	**94.2**

Table 1: Image-Text Matching on MSCOCO 1K and Flickr30K.

- CC120K. We randomly sample a subset from the real-world dataset Conceptual Captions (Sharma et al. 2018). This dataset is harvested from the Internet, with about 3%-20% incorrect image-text pairs. CC120K contains 120, 851 with a single caption per image. In our experiment, we use 118,851 images for training, 1,000 images for validation, and 1,000 images for testing.

The widely-used metric Recall@K (R@K) is used to evaluate the performance of image-text matching with K=1, 5, and 10. The variance (var) of R@1 at different noise ratios is used to evaluate the approaches' performance stability, with lower var indicating higher stability.

Implementation Details. NPC can enhance noise resistance and stability in various cross-modal matching models. In this paper, the CLIP (Radford et al. 2021) with ViT-B/32 is implemented as a baseline. Both baseline and NPC are trained on a single RTX 3090 GPU optimized by AdamW (Loshchilov and Hutter 2019). We start training CLIP and NPC with learning rates $5e-7$ and $2e-7$ with a weight decay of 0.2. In all experiments, we train the model for 5 epochs with a mini-batch size of 256, and the hyperparameter τ is set to 0.99.

Comparison with State of the Arts

Quantitative Comparison. To illustrate the effectiveness, we compare NPC with various approaches, including general cross-modal matching methods SCAN (Lee et al. 2018), SAF (Diao et al. 2021), noise-robust learning methods NCR (Huang et al. 2021), DECL (Qin et al. 2022), BiCro (Yang et al. 2023), and CLIP with fine-tuning (Radford et al. 2021). It is worth noting that CLIP is the baseline of our method. The results are shown in Table 1.

It shows that NPC significantly outperforms all methods across all noise ratios. Notably, on Flickr30K with 60% noise ratio, NPC outperforms the current state-of-the-art approach BiCro with a large R@1 performance gap. To be specific, the R@1 performance of NPC is 15.4% higher than BiCro on image-text matching (i2t), as well as 16.9% higher than BiCro on text-to-image matching (t2i). Compared to the baseline CLIP, NPC has achieved immense improvement in all metrics and benchmarks. Furthermore, as the noise ratio increases, the performance gap between NPC and baseline becomes larger. For instance, on the MSCOCO 1K set, when the noise ratio ranges from 0% to 60%, the R@1 performance gap between NPC and baseline separately increases from 2.3% to 11.2% on i2t, and 3.3% to 13.4% on t2i. This

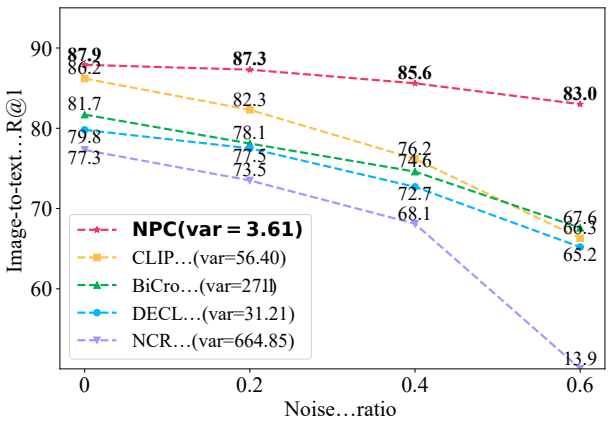


Figure 4: Variation and variance (var) of matching performance at different noise ratio.

method	image-to-text			text-to-image		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP	68.8	87.0	92.9	67.8	86.4	90.9
NPC	**71.1**	**92.0**	**96.2**	**73.0**	**90.5**	**94.8**

Table 2: Comparison with baseline on CC120K.

phenomenon is powerful to prove the effectiveness of NPC on robust learning.

Stability Comparison. To further explore the superiority of NPC on stable learning, we illustrate the R@1 change curves of different methods under different noise ratios in Fig. 4. We can observe that NPC outperforms all other methods in all noise ratios. Meanwhile, as the noise ratio increases, the performance decline of NPC is significantly smaller than that of other methods. Furthermore, we calculate the variance of each method on different noise ratios to quantify the stability of the methods. NPC shows remarkable stability with only 3.61% variance, outperforming all other methods with a huge gap. Compared to the baseline CLIP, NPC yields a large drop on var of 52.79%. The large decrease in variance indicates the performance stability is significantly improved by NPC.

Comparison with ViT-B/32 Backbone Methods

In Table 2, we compare the NPC with baseline on the CC120K which is with real noisy correspondences. From the results, our proposed method outperforms the baseline by a considerable margin in terms of all metrics. Specifically, NPC is 2.3% and 5.2% higher than CLIP on i2t and t2i R@1, respectively.

For a fair comparison, we also compare the NPC to the methods with the same CLIP ViT-B/32-based backbone, including VSE∞ (Chen et al. 2021), PCME (Chun et al. 2021), PCME++ (Chun 2023), and PAU (Li et al. 2023). The results on noise-free MSCOCO 5K are shown in Table 3. It demonstrates that NPC consistently outperforms other methods in all metrics. Besides, we also report the average R@1 of image-to-text and text-to-image of MSCOCO 1K and 5K

method	image-to-text			text-to-image		
	R@1	R@5	R@10	R@1	R@5	R@10
VSE∞	60.2	85.4	92.2	46.9	75.5	84.8
PCME	59.9	85.8	92.3	46.1	75.0	84.6
PCME++	61.8	87.0	93.0	47.9	76.5	85.4
PAU	63.6	85.2	92.2	46.8	74.4	83.7
CLIP	62.2	84.6	90.9	45.1	72.3	81.8
NPC	**65.4**	**87.3**	**93.1**	**48.5**	**75.4**	**84.4**

Table 3: Comparison of methods with ViT-B/32 backbone on noise-free MSCOCO 5K.

noise	method	1K R@1	5K R@1	1K RSUM
20%	VSE∞	72.0	51.4	520.2
	PCME	69.9	48.1	519.3
	PCME++	70.8	49.5	522.4
	PAU	71.4	51.7	521.5
	CLIP	66.8	47.2	507.2
	NPC	**73.1**	**53.8**	**529.8**
50%	VSE	38.5	18.4	390.5
	PCME	65.8	43.0	505.7
	PCME++	65.7	44.0	503.9
	PAU	69.3	49.6	513.4
	CLIP	60.9	41.4	486
	NPC	**71.3**	**51.9**	**523.4**

Table 4: Comparison of methods with ViT-B/32 backbone on noisy MSCOCO.

in Table 4 at different noise ratios. Meanwhile, the sum of R@1, R@5, and R@10 on both i2t and t2i on MSCOCO 1K is also reported. As the noise ratio increases, NPC outperforms others by larger margins, surpassing the second best model PAU by 2.0% at 20% noise ratio, while 2.3% at 50% noise ratio for 5K R@1. All these experiments effectively demonstrate the effectiveness and superiority of NPC.

Ablation Study

Analysis on w and $\mathcal{L}_{MB}$. According to Eq. 11, there are two important components of confidence weight w and memory bank loss $\mathcal{L}_{MB}$ in the re-training process. To explore the effect of each component, we exhaustively ablate them in Flickr30K with three noise ratios. The results are shown in Table 5.

We observe that both w and $\mathcal{L}_{MB}$ obtain significant performance improvements in different noise ratios. They bring almost the same improvements for NPC compared with the baseline. Specifically, training with 60% noise, the ablative NPC exceeds the baseline by 11.8% and 6.95% on average R@1 of image-to-text and text-to-image, indicating that w and $\mathcal{L}_{MB}$ have independent effect of anti-noise. Moreover, the full version NPC outperforms by a much larger margin than the baseline, indicating that both components can complement each other and collaborate to achieve robust learning. The reason why w and $\mathcal{L}_{MB}$ can achieve robust learning is that the confidence weight w mitigates the degree of negative impact from the noisy sample to the model, and the memory bank loss $\mathcal{L}_{MB}$ can provide correct correspondences for these noisy samples.

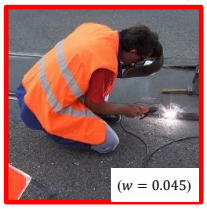
1. Bird 's eye view of a city worker walking on the street. (w = 0.037)
2. A man windsurfs with a group of other windsurfers. (w = 1)
3. A man is using a chisel to do wood work on a wooden pole. (w = 0.042)
4. The man is on a ski boat in the water. (w = 1)
5. A man windsurfing. (w = 1)

1. An ethnic woman is sewing some sort of blue and gold material with elaborate designs. (w = 0.042)
2. A woman in a purple sweatshirt with a large ivory hat is painting a scene of a boat launch on her canvas, affixed to an easel. (w = 0.046)
3. Two brothers preparing the family dinner. (w = 1)
4. Two boys with blue and green on cooking. (w = 1)
5. Many marathon runners on a stretch of path in a wintry park. (w = 0.102)

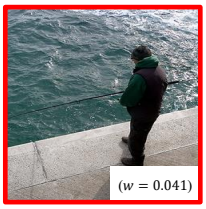
(a) The image and its 5 annotated captions. The noise captions are in red. The correct captions are in green.

A man with short hair is apprehended and restrained on a street.

Two boys are walking down a brick sidewalk.



People are sitting on a park bench while there are 3 pictures visible in the foreground.



A guy wearing blue in a hole.

(b) The caption and its noisy correspondent image.

Figure 5: Some noisy correspondences in Flickr30K training set under 40% noise. The average confidence weight (w) in all epochs is shown for examples. The w of correctly matched pairs are obviously larger than noisy pairs.

noise	threshold τ	image-to-text			text-to-image		
		R@1	R@5	R@10	R@1	R@5	R@10
0%	0.5	87.2	**98.1**	99.2	74.5	**93.7**	96.9
	0.7	87.6	97.9	**99.4**	74.9	93.5	97.1
	0.99	**87.9**	**98.1**	**99.4**	**75.0**	**93.7**	**97.2**
60%	0.5	78.3	94.2	96.7	59.2	82.6	88.8
	0.7	82.2	**95.9**	98.3	67.8	89.4	**94.2**
	0.99	**83.1**	**95.9**	98.6	**68.1**	**89.6**	**94.2**

Table 5: Ablation study of threshold τ on Flickr30k.

noise	method		image-to-text			text-to-image		
	w	$\mathcal{L}_{MB}$	R@1	R@5	R@10	R@1	R@5	R@10
20%	✓	✓	**87.3**	**97.5**	**98.8**	**72.9**	**92.1**	**95.8**
	✓		85.3	97.3	**98.8**	71.8	91.3	95.2
		✓	85.4	97.2	98.6	71.9	91.4	95.2
			82.3	95.5	98.3	66.0	88.5	93.5
40%	✓	✓	**85.6**	**97.5**	**98.4**	**71.3**	**91.3**	**95.3**
	✓		79.9	95.5	97.7	62.4	85.5	91.1
		✓	79.0	95.0	97.5	62.3	85.2	91.1
			76.2	93.3	96.5	59.4	85.0	90.9
60%	✓	✓	**83.0**	**95.9**	**98.6**	**68.1**	**89.6**	**94.2**
	✓		78.2	93.5	96.8	59.0	82.5	88.4
		✓	78.0	93.9	96.6	59.1	82.3	88.7
			66.3	87.3	93.0	52.1	78.8	87.4

Table 6: Ablation studies for w and $\mathcal{L}_{MB}$ on Flickr30K.

Analysis on hyperparameter τ. τ is a very important parameter, which can control the clean degree of clean set D_c in Eq. 5 and memory bank MB. A smaller value of τ leads to a larger scale of D_c, potentially containing more noise pairs. The purity of D_c directly impacts the quality of MB, which in turn influences the model's matching performance. To explore the impact of the selection threshold τ on the model, we report the matching performance with different τ on Flickr30K with 0% and 60% noise ratios in Table 6, respectively. The results show that when training with 0% noise, the impact of varying τ on performance reduction is not noticeable. However, in the case of training with 60% noise, performance drops by 4.8% and 7.0% on R@1 when τ changes from 0.99 to 0.5. It implies that a rigorous selection of D_c is necessary to establish a trustworthy MB.

Visualization

To illustrate the effectiveness of NPC, we showcase examples from Flickr30K in Fig. 5. The average confidence weight (w) for each pair across five epochs is depicted. Noisy pairs consistently exhibit notably low w values. Especially in Fig. 5 (a), there is a very obvious contrast between the w of the same image with correct annotations and noisy annotations. That is to say, with the support of MB, NPC effectively differentiates between clean and noisy correspondences. It also avoids model learning errors by assigning a small w to the noisy correspondence.

Conclusion

This paper studies a novel challenge of maintaining stable performance for the noise-robust learning model as noise increases. To tackle this, a novel approach NPC is proposed. We introduce a novel NP paradigm to estimate per-sample negative impact before it is learned by the model. To obtain the negative impact, the memory bank of the training set is constructed by strict selection. To mitigate negative impact on the model, each sample is assigned a confidence weight based on the memory bank. Extensive experiments indicate the effectiveness of each component in our method. The NPC achieves notable enhancement in matching accuracy and performance stability compared to the state-of-the-art approach on both noise and noise-free datasets.

Acknowledgement. This work is partially supported by National Natural Science Foundation of China under Grants (62176188) and the Key Research and Development Program of Hubei Province (2021BAD175)

References

Arpit, D.; Jastrzebski, S.; Ballas, N.; Krueger, D.; Bengio, E.; Kanwal, M. S.; Maharaj, T.; Fischer, A.; Courville, A. C.; Bengio, Y.; and Lacoste-Julien, S. 2017. A Closer Look at Memorization in Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning, (ICML-17)*, volume 70 of *Proceedings of Machine Learning Research*, 233–242. Sydney, NSW, Australia: PMLR.

Chen, J.; Hu, H.; Wu, H.; Jiang, Y.; and Wang, C. 2021. Learning the Best Pooling Strategy for Visual Semantic Embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR-21)*, 15789–15798. virtual: IEEE.

Chun, S. 2023. Improved Probabilistic Image-Text Representations. arXiv:2305.18171.

Chun, S.; Oh, S. J.; de Rezende, R. S.; Kalantidis, Y.; and Larlus, D. 2021. Probabilistic Embeddings for Cross-Modal Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR-21)*, 8415–8424. virtual: IEEE.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT-19)*, 4171–4186. Minneapolis, MN, USA: Association for Computational Linguistics.

Diao, H.; Zhang, Y.; Ma, L.; and Lu, H. 2021. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI conference on artificial intelligence (AAAI-21)*, 1218–1226. Palo Alto, California: AAAI Press.

Ding, M.; Yang, Z.; Hong, W.; Zheng, W.; Zhou, C.; Yin, D.; Lin, J.; Zou, X.; Shao, Z.; Yang, H.; and Tang, J. 2021. CogView: Mastering Text-to-Image Generation via Transformers. In *Advances in Neural Information Processing Systems (NeurIPS-21)*, 19822–19835. virtual.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.

Faghri, F.; Fleet, D. J.; Kiros, J. R.; and Fidler, S. 2018. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In *British Machine Vision Conference 2018 (BMVC-18)*, 12. Newcastle, UK: BMVA Press.

Han, H.; Miao, K.; Zheng, Q.; and Luo, M. 2023. Noisy Correspondence Learning with Meta Similarity Correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR-23)*, 7517–7526. IEEE.

Huang, Z.; Niu, G.; Liu, X.; Ding, W.; Xiao, X.; Wu, H.; and Peng, X. 2021. Learning with noisy correspondence for cross-modal matching. In *Advances in Neural Information Processing Systems (NeurIPS-21)*, volume 34, 29406–29419. virtual.

Jiang, D.; and Ye, M. 2023. Cross-Modal Implicit Relation Reasoning and Aligning for Text-to-Image Person Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR-23)*, 2787–2797.

Lee, K.-H.; Chen, X.; Hua, G.; Hu, H.; and He, X. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV-18)*, 201–216. Munich, Germany: Springer.

Lei, S. W.; Gao, D.; Wu, J. Z.; Wang, Y.; Liu, W.; Zhang, M.; and Shou, M. Z. 2023. Symbolic replay: Scene graph as prompt for continual learning on vqa task. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-23)*, 1250–1259. Washington, DC, USA: AAAI Press.

Li, H.; Song, J.; Gao, L.; Zeng, P.; Zhang, H.; and Li, G. 2022. A Differentiable Semantic Metric Approximation in Probabilistic Embedding for Cross-Modal Retrieval. In *Advances in Neural Information Processing Systems (NeurIPS-22)*, volume 35, 11934–11946.

Li, H.; Song, J.; Gao, L.; Zhu, X.; and Shen, H. T. 2023. Prototype-based Aleatoric Uncertainty Quantification for Cross-modal Retrieval. In *NeurIPS*.

Li, J.; Socher, R.; and Hoi, S. C. H. 2020. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. arXiv:2002.07394.

Li, K.; Zhang, Y.; Li, K.; Li, Y.; and Fu, Y. 2019a. Visual semantic reasoning for image-text matching. In *2019 IEEE/CVF International Conference on Computer Vision, (ICCV-19)*, 4654–4662. Seoul, Korea (South): IEEE.

Li, S.; Tao, Z.; Li, K.; and Fu, Y. 2019b. Visual to text: Survey of image and video captioning. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 3: 297–312.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European conference on computer vision (ECCV-14)*, 740–755. Zurich, Switzerland: Springer.

Lin, Y.; Xie, Y.; Chen, D.; Xu, Y.; Zhu, C.; and Yuan, L. 2022. Revive: Regional visual representation matters in knowledge-based visual question answering. In *Advances in Neural Information Processing Systems (NeurIPS-22)*, volume 35, 10560–10571.

Liu, S.; Niles-Weed, J.; Razavian, N.; and Fernandez-Granda, C. 2020. Early-Learning Regularization Prevents Memorization of Noisy Labels. In *Advances in Neural Information Processing Systems (NeurIPS-20)*, volume 33, 20331–20342.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations (ICLR-19)*. New Orleans, LA, USA: OpenReview.net.

Permuter, H.; Francos, J.; and Jermyn, I. 2006. A study of Gaussian mixture models of color and texture features for image classification and segmentation. *Pattern recognition*, 39(4): 695–706.

Qian, S.; Xue, D.; Zhang, H.; Fang, Q.; and Xu, C. 2021. Dual adversarial graph neural networks for multi-label cross-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-21)*, 2440–2448. Palo Alto, California: AAAI Press.

Qin, Y.; Peng, D.; Peng, X.; Wang, X.; and Hu, P. 2022. Deep evidential learning with noisy correspondence for cross-modal retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM-22)*, 4948–4956. New York, NY, United States: Association for Computing Machinery.

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 *Proceedings of the 38th International Conference on Machine Learning ICML-21*, 8748–8763. Virtual Event: PMLR.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 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 (ACL-18)*, 2556–2565. Melbourne, Australia: Association for Computational Linguistics.

Song, Y.; and Soleymani, M. 2019. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR-19)*, 1979–1988. Long Beach, CA, USA: IEEE.

Stefanini, M.; Cornia, M.; Baraldi, L.; Cascianelli, S.; Fiameni, G.; and Cucchiara, R. 2022. From show to tell: A survey on deep learning-based image captioning. *IEEE transactions on pattern analysis and machine intelligence*, 45: 539–559.

Wang, L.; Li, Y.; Huang, J.; and Lazebnik, S. 2018. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 394–407.

Wang, N.; Xie, J.; Luo, H.; Cheng, Q.; Wu, J.; Jia, M.; and Li, L. 2023. Efficient Image Captioning for Edge Devices. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-23)*, 2608–2616. Washington, DC, USA: AAAI Press.

Xia, X.; Liu, T.; Han, B.; Gong, C.; Wang, N.; Ge, Z.; and Chang, Y. 2021. Robust early-learning: Hindering the memorization of noisy labels. In *9th International Conference on Learning Representations, (ICLR-21)*. Virtual Event, Austria: OpenReview.net.

Yang, S.; Xu, Z.; Wang, K.; You, Y.; Yao, H.; Liu, T.; and Xu, M. 2023. BiCro: Noisy Correspondence Rectification for Multi-modality Data via Bi-directional Cross-modal Similarity Consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR-23)*, 19883–19892. IEEE.

Ye, M.; Li, H.; Du, B.; Shen, J.; Shao, L.; and Hoi, S. C. H. 2022. Collaborative Refining for Person Re-Identification With Label Noise. *IEEE Trans. Image Process.*, 31: 379–391.

Ye, M.; and Yuen, P. C. 2020. PurifyNet: A Robust Person Re-Identification Model With Noisy Labels. *IEEE Trans. Inf. Forensics Secur.*, 15: 2655–2666.

You, H.; Guo, M.; Wang, Z.; Chang, K.-W.; Baldridge, J.; and Yu, J. 2023. CoBIT: A Contrastive Bi-directional Image-Text Generation Model. arXiv:2303.13455.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2: 67–78.

Zhang, X.; Niu, X.; Fournier-Viger, P.; and Dai, X. 2023. Image-text Retrieval via Preserving Main Semantics of Vision. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, 1967–1972.

Zhou, Y.; Zhang, R.; Gu, J.; Tensmeyer, C.; Yu, T.; Chen, C.; Xu, J.; and Sun, T. 2022. Tigan: Text-based interactive image generation and manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-22)*, 3580–3588. Virtual Event: AAAI Press.