Noise-Aware Image Captioning with Progressively Exploring Mismatched Words

Zhongtian Fu[1*], **Kefei Song**[1*], **Luping Zhou**[2], **Yang Yang**[1†]

[1]Nanjing University of Science and Technology, Nanjing 210094, China.
[2]The University of Sydney, Sydney 2052, Australia.
ztfu@njust.edu.cn, kfsong@njust.edu.cn, luping.zhou@sydney.edu.au, yyang@njust.edu.cn

Abstract

Image captioning aims to automatically generate captions for images by learning a cross-modal generator from vision to language. The large amount of image-text pairs required for training is usually sourced from the internet due to the manual cost, which brings the noise with mismatched relevance that affects the learning process. Unlike traditional noisy label learning, the key challenge in processing noisy image-text pairs is to finely identify the mismatched words to make the most use of trustworthy information in the text, rather than coarsely weighing the entire examples. To tackle this challenge, we propose a Noise-aware Image Captioning method (NIC) to adaptively mitigate the erroneous guidance from noise by progressively exploring mismatched words. Specifically, NIC first identifies mismatched words by quantifying word-label reliability from two aspects: 1) inter-modal representativeness, which measures the significance of the current word by assessing cross-modal correlation via prediction certainty; 2) intra-modal informativeness, which amplifies the effect of current prediction by combining the quality of subsequent word generation. During optimization, NIC constructs the pseudo-word-labels considering the reliability of the origin word-labels and model convergence to periodically coordinate mismatched words. As a result, NIC can effectively exploit both clean and noisy image-text pairs to learn a more robust mapping function. Extensive experiments conducted on the MS-COCO and Conceptual Caption datasets validate the effectiveness of our method in various noisy scenarios.

1 Introduction

In reality, people perceive information from various modalities like vision, hearing, and smell. With the development of artificial intelligence, especially deep learning, multi-modal learning has received increasing attention (Hossain et al. 2019; Yang et al. 2020, 2021). Image captioning is one of the important downstream tasks, which aims to automatically generate meaningful descriptions for images and has a wide range of applications in daily lives (Hossain et al. 2019; Baltrusaitis, Ahuja, and Morency 2019; Yang et al. 2023). For example, E-commerce benefits from quick product image descriptions, and visually impaired people use captioning models to experience the visual world.

Indeed, image captioning research focuses on learning generators between heterogeneous image-text modalities. This involves comprehending the given image through computer vision techniques and generating corresponding descriptions using natural language processing. Initially, researchers explored the encoder-decoder architecture (Yang et al. 2019a; Zhang et al. 2019; Yang et al. 2019c) with CNNs (Albawi, Mohammed, and Al-Zawi 2017) as image encoders and LSTM (Greff et al. 2017) as text decoders (Vinyals et al. 2015). To consider local and global features simultaneously, (Huang et al. 2019) used image regions to decode image segmentations sequentially to words, adding attention mechanisms to focus on specific image regions during decoding. With the advancement in multi-head self-attention-based Transformers, they improved the modeling of intra-modal and inter-modal relationships, leading to better captions (Cornia et al. 2020; Zhang et al. 2021; Yang et al. 2022). Notably, effective deep captioning requires large-scale image-text pairs, such as CLIP using 400 million image-text pairs for training (Radford et al. 2021), typically collected from the internet to reduce labor costs. However, an issue is the presence of noisy image-text pairs, where mismatches significantly degrade model performance. Researchers have attempted to address this by explicitly weighting image-text pairs (Huang et al. 2019, 2021), with smaller weights indicating more noise. Yet, noisy samples in image captioning scenarios are fine-grained, differing from coarse-grained label noise in noisy label learning. Figure 1 illustrates different noise levels: matched, partially matched, and mismatched. The degree of noise is a continuous value from 0 to 1, where 0 denotes mismatched or useless pairs, and 1 means matched or clean pairs. Partially matched pairs still contain useful word-region pairs, with lower noise degrees indicating more useful pairs. Thus, the noisy image-text pair can be considered the ground truth sentence containing incorrect words.

Along this line, in this paper, we mitigate the influence of noisy image-text pairs by fine-grained exploring mismatched words and progressively conducting more reliable pseudo-word-labels. Naturally, we propose a Noisy-aware Image Captioning method (NIC), which learns image-text pairs with the newly designed reliability evaluation mecha-

*These authors contributed equally.
†Corresponding author.

(a) (b) (c)

Figure 1: Example of different levels of noise. Image-text pairs from the real world usually contain noise in varying degrees, e.g., (a), (b), and (c) respectively show the matched, partially matched, and mismatched samples. Valid object words in the caption have been bolded, where the blue ones map relevant regions in the image, while the red ones do not.

nism at the word level, comprehensively considering both the inter-modal and intra-modal influences. Specifically, NIC learns the reliability weight of word-labels from two aspects: Inter-modal representativeness focuses on information certainty of the current word's prediction, which reflects the relevance of the current word to image contents; Intra-modal informativeness further inspects the quality of subsequent word generation based on previous prediction. During iterative model optimization, we construct pseudo-word-labels based on word-label reliability and model convergence to coordinate mismatched words, enhancing robustness and accuracy in the supervised constraints. Moreover, as a generalized framework, NIC can be applied to any existing image captioning models, and several typical approaches (Huang et al. 2019; Zhang et al. 2021; Luo et al. 2021; Wang, Xu, and Sun 2022) are adopted for validation.

In summary, the contributions of this paper can be summarized as follows:

- We explore the noisy image captioning in a fine-grained manner, which carefully measures the word-labels considering both the inter-modal and intra-modal influences to identify mismatched words;
- We design a noise-aware image captioning approach, which can progressively conduct pseudo-word-labels to adaptively coordinate mismatched words for robustness;
- In experiments, our method improves the performance in all noisy scenes, which validates its effectiveness.

2 Related Work

2.1 Image Captioning

Image captioning aims to automatically generate a sentence to describe the visual content of a given image. Early works designed a template-based approach (Farhadi et al. 2010; Li et al. 2011), which manually generated a caption template and filled it with objects, attributes, and relationships detected from the images. However, due to the premade nature of the template, only descriptions of a specific length may be produced, which limits flexibility. With the development of deep learning, modern image captioning methods have achieved breaking advances. Inspired by neural machine translation (Cho et al. 2014), researchers have explored the encoder-decoder architecture. (Vinyals et al. 2015) used CNN to encode the images and LSTM to generate sentences. (Karpathy and Fei-Fei 2017) used bi-directional RNN as decoders and aligned two modalities. To align each word with relevant visual content, (You et al. 2016) designed a semantic attention module that measures the semantic importance of visual objects. (Huang et al. 2019) added the attention on the attention module to refine attention results. Encouraged by the success of transformers in refining sequence modeling (Vaswani et al. 2017), (Herdade et al. 2019) used the self-attention module to model the relationships between image regions. (Wang, Xu, and Sun 2022) designed a refined encoder to capture the internal relationships between the grid features. To achieve better performance, the above methods usually require a large amount of image-text pairs as training data. However, the texts may be noisy due to various reasons, such as limited crowdsourcing, low-quality labeled web-crawled images, etc.

2.2 Noisy Label Learning

Learning from noisy labels is an important task as the presence of noisy labels can significantly degrade the model's generalization performance. Various methods have been proposed to train robust models under noisy labels, which can be roughly categorized into three groups: 1) Robust architecture. For instance, (Chen and Gupta 2015) added a noise adaptation layer on top of the softmax layer; (Cheng et al. 2020) proposed to use an edge information network for model architecture. 2) Robust loss function. For example, (Wang et al. 2019) used weighted sums of cross-entropy loss and reverse cross-entropy loss to construct robust loss functions; (Lyu and Tsang 2020) proposed a curriculum loss to adaptively select samples for training. 3) Loss adjustment. For example, (Reed et al. 2015) proposed to use a combination of raw labels and predicted labels to calculate the loss; (Wang et al. 2021) utilized the entropy of predicted labels to progressively correct semantic classes. However, these methods are primarily designed for image classification tasks, and directly transferring them to cross-modal tasks can not yield optimal results. To better solve the noisy problem in cross-modal scenes, (Huang et al. 2021) designed a triplet loss with soft edges, and (Qin et al. 2022) proposed a robust dynamic hinge loss to capture and learn from uncertainty in cross-modal retrieval tasks. (Huang et al. 2019; Luo et al. 2021) directly set the same smoothing parameters for all examples in the image captioning task. (Li et al. 2021) introduced momentum distillation for learning from pseudo-targets, and (Kang et al. 2023) utilized pre-trained CLIP to compute image-text similarity, serving as a training control signal for guiding the model in learning diverse alignment levels, both drawing from prior knowledge of the large-scale models. However, these methods only processed the holistic examples at a coarse level or relied on additional prior knowledge. In contrast, we explore mismatched words at a fine-grained level by considering inter-modal and intra-modal information from the data itself and provide flexible loss constraints.

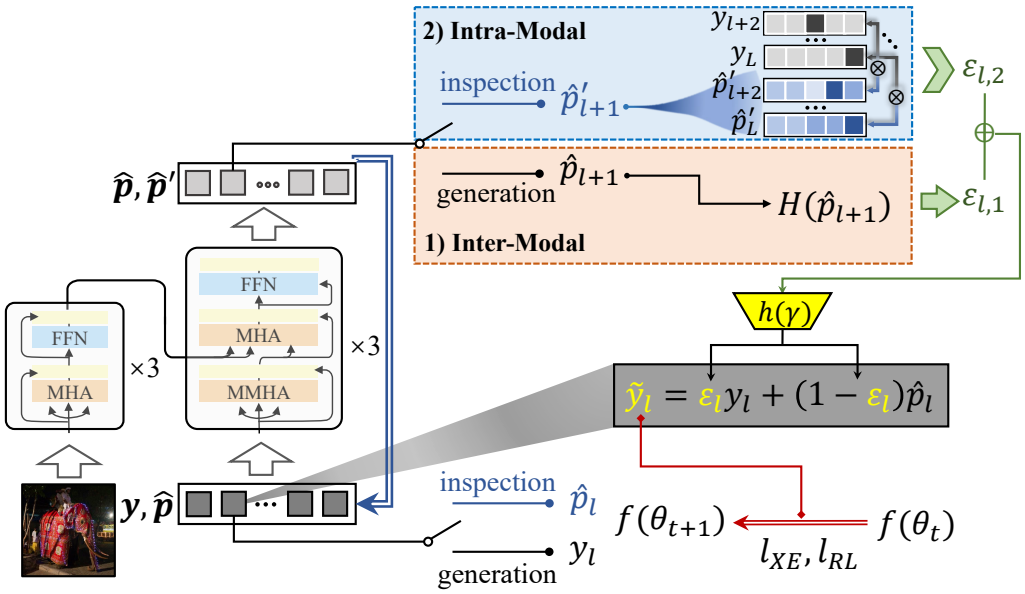


Figure 2: Framework of the proposed NIC. NIC first distinguishes mismatched words from two aspects: 1) Inter-Modal Representativeness, which considers the entropy of predicted word $\hat{p}_{l+1}$ as $\epsilon_{l,1}$ to measure the relevance between the current word and the image; 2) Intra-Modal Discriminativeness, which calculates the cross-entropy of context ($\otimes$) to obtain $\epsilon_{l,2}$ to reflect the stability of sequence generation. Meanwhile, to further exploit the mismatched words, NIC combines the reliability weight (ϵ_l) with the model convergence ($h(\gamma)$) to construct the pseudo-word-label $\tilde{y}_l$ for model updating.

3 Proposed Method

Figure 2 illustrates an overview of NIC. We add an extra inspection process for identifying word-labels compared with traditional captioning methods. On the one hand, inter-modal semantic relevancy is quantified by measuring the certainty of word-predictions during the generation process. On the other hand, additional predictions, which source from feeding the original prediction into the decoder(i.e., the blue arrow), are considered to enlarge the gap between clean and noisy words inspired by error accumulation. Combined with model convergence $h(\gamma)$, the aggregated reliability weight ϵ_l reflects the reliability of each word-label, which is further used to construct more reliable pseudo word-label for more robust parameter updating (i.e., the red arrow).

Basic Notations For given image-text pairs $D = \{\boldsymbol{i}_o, \boldsymbol{s}_o\}_{o=1}^{O}$ in noisy scenarios, $\boldsymbol{i}_o$, $\boldsymbol{s}_o$ denotes the $o-$th image and text respectively, and O is the number of examples. There exist noisy pairs, where contents of $\boldsymbol{i}_o$ mismatch or partially matches $\boldsymbol{s}_o$, providing biased guidance information for training and leading to a drop in performance.

3.1 Cross-modal Generator

NIC is a versatile framework that can be applied with various state-of-the-art cross-modal models. Considering the effectiveness, here NIC takes the well-known Transformer as an example, essentially an Encoder-Decoder architecture. Standard Transformer model (Vaswani et al. 2017) is usually applied to NLP tasks at early stages and further extended to the study of image captioning rapidly by virtue of its powerful learning ability based on multi-head cross-attention mechanism. The encoder aims to model the relationships between image regions to obtain semantically richer image region features, which contains two sub-layers, a multi-head attention module (MHA), and a position-wise feed-forward network (FFN). To align the visual-language modality, the decoder first takes the caption as input and uses the mask matrix to compute the multi-head attention at each moment in parallel to obtain the language token features $\boldsymbol{e}^S$. Next, $\boldsymbol{e}^S$ projected as query along with visual area features $\boldsymbol{e}^I$ from encoder projected as key and value are fed to the MHA and then FFN for the generation of predictions.

Based on the basic model structure, model learning is generally divided into training and fine-tuning. During training, optimization aims to minimize the cross-entropy loss:

$$\ell_{XE}(\theta) = -\sum_{l=1}^{L} \log p_\theta(y_l | y_{1:l-1}), \tag{1}$$

where $y_{1:L}$ denotes the ground truth, and L represents the number of tokens in the caption. The parameters θ of the network define a policy p_θ. Afterward, during fine-tuning, researchers optimize the non-differentiable metrics with Self-Critical Sequence Training (Rennie et al. 2017) based on the best model obtained in the previous stage:

$$\ell_{RL}(\theta) = -\mathbb{E}_{y_{1:L} p_\theta}\left[r\left(y_{1:L}\right)\right]. \tag{2}$$

The reward $r(\cdot)$ is a sentence-level metric for the generated sentence and the ground truth, which is always represented by the score of captioning metric (e.g., CIDEr (Vedantam, Zitnick, and Parikh 2015)).

3.2 Noise-Aware Identification

Inspired by limited gains in performance from weighting image-text pairs, we emphasize fine-grained mechanisms from both inter-modal and intra-modal perspectives.

Inter-Modal Representativeness In image captioning, the decoder predicts the current word based on preceding text and visual features, essentially performing a classification task within the vocabulary. Combined with corresponding cross-modal attention and context, the decoder yields more confident predictions, which is not the case for noisy words nevertheless. The probability distribution of predictions is influenced by image-text correspondence, and following (Wang et al. 2021), the prediction appears high confidence when the entropy remains low. Thus, entropy can serve to quantify the quality of generation based on visual-textual relevance. For target word-label y_l, $\epsilon_{l,1}$ is introduced as the inter-modal representativeness assessment:

$$\epsilon_{l,1} = 1 - Norm\left(H\left(\hat{p}_{l+1}\right)\right), \tag{3}$$

where $\hat{p}_{l+1}$ denotes model prediction based on y_l. $H(\cdot)$ denotes calculation of entropy. As the higher the entropy, the higher the uncertainty, $\epsilon_{l,1}$ about noisy word-labels tend to have lower values compared with clean ones, which reflect the image-word correlation from prediction confidence.

Intra-Modal Discriminativeness Unlike image classification, image captioning is sequential, where contextual information affects subsequent word generation. The quality of prefix words as guidance largely determines subsequent word generation. Therefore, we propose a Subsequent Prediction Inspection strategy to judge whether the target word-label is confident and valid. As Figure 3 takes word-label y_l as an example, traditional methods only get prediction sequence with given word-labels, i.e., $\hat{\boldsymbol{p}} = \hat{p}_{1:L}$, which is later directed to approach ground truth. In contrast, after y_l has been absorbed, NIC successively feeds word-predictions in $\hat{\boldsymbol{p}}$ into the decoder as the mock label each moment to generate additional subsequent word-predictions $\hat{\boldsymbol{p}}_{\boldsymbol{l}}{}'$ for y_l:

$$\begin{aligned} \hat{\boldsymbol{p}}_{\boldsymbol{l}}{}' &= \hat{p}'_{l+2:L} = \left\{\hat{p}'_{\hat{l}}\right\}_{\hat{l}=l+2}^{L}, \\ \hat{p}'_{\hat{l}} &= Decoder(\boldsymbol{e}^I, y_{<l+1} \cup \{\hat{p}_{l+1}\} \cup \hat{p}'_{l+2:\hat{l}-1}). \end{aligned} \tag{4}$$

Quality of $\hat{\boldsymbol{p}}_{\boldsymbol{l}}{}'$ reveals rationality of given word-label y_l in that the mock label (original word-prediction) based on a noisy word-label always leads to worse subsequent word-predictions. Notably, additional predictions gradually lose more semantic information when the target word-label is noisy, resulting in error accumulation. Thus, $\hat{\boldsymbol{p}}_{\boldsymbol{l}}{}'$ with noisy original predictions as the source corresponds to a more exaggerated loss value compared to clean words. Given this, $\epsilon_{l,2}$ is introduced as the intra-modal discriminativeness assessment considering the cross-entropy loss brought by $\hat{\boldsymbol{p}}_{\boldsymbol{l}}{}'$:

$$\epsilon_{l,2} = 1 - Norm\left(\frac{1}{L-(l+2)}\ell_{XE}\left(y_{l+2:L}, \hat{p}'_{l+2:L}\right)\right), \tag{5}$$

where $y_{l+2:L}$ is the annotated word-labels, and $\hat{p}'_{l+2:L}$ is the word-predictions with a mixture of word-labels and original word-predictions as the input source of the decoder.

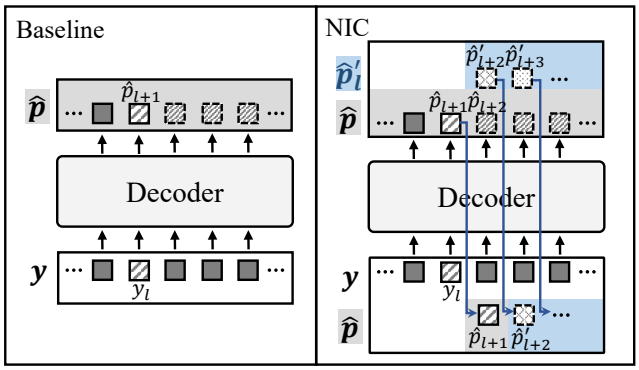


Figure 3: Comparison chart about Baseline and NIC on decoding stage with noisy word-label y_l as an example. Apart from normal prediction sequence $\hat{\boldsymbol{p}}$, NIC generates additional predictions $\hat{\boldsymbol{p}}_{\boldsymbol{l}}{}' = \hat{p}'_{l+2:L}$ with regard to each y_l based on previous word-predictions.

3.3 Progressive Coordination

Since traditional image captioning methods ignore the noise in labels, the wrong visual-to-language mapping relationship is absorbed undesirably, i.e. Eq. 1 and Eq. 2, which leads to a huge drop in performance. From a label-fitting standpoint, we conduct pseudo-word-labels via a progressive coordination strategy, where the model fits reliable and clean word-labels as usual, but gradually places more trust in its predictions for noisy word-labels during the training process. As mentioned above, the reliability weight of each word-label can be defined by $\epsilon_{l,1}$ and $\epsilon_{l,2}$, e.g., $\epsilon_{l,1} + \epsilon_{l,2}$. Considering that the model fits the noise in varying degrees as the training progresses to different stages, model convergence (i.e., iterations) will be taken into consideration:

$$\begin{aligned} h(\gamma) &= 1 - \frac{1}{1+\exp(-(\frac{\gamma}{\Gamma} - C) * \tau)}, \\ \epsilon_l &= h(\gamma) * (\epsilon_{l,1} + \epsilon_{l,2}), \end{aligned} \tag{6}$$

where γ and Γ respectively represent the current iteration and total iterations; C and τ are the hyperparameters to adjust the curvature. ϵ_l denotes the reliability, whose value is higher for the clean word compared with the noisy one. Then, we fuse each word-label and its corresponding model prediction to construct a more reliable pseudo-word label based on ϵ_l and modify the constraint:

$$\begin{aligned} \tilde{y}_l &= \epsilon_l y_l + (1-\epsilon_l)\hat{p}_l, \\ \ell'_{XE}(\theta) &= -\sum_{l=1}^{L} \log p_\theta(\tilde{y}_l | y_{1:l-1}), \end{aligned} \tag{7}$$

where y_l is the ground truth, $\hat{p}_l$ is the prediction relied on $y_{<l}$ and $\tilde{y}_l$ is the constructed pseudo label eventually. In the fine-tuning stage, $\ell_{RL}(\theta)$ optimizes the model parameters using CIDEr scores. Notably, CIDEr averages cosine similarity between predicted sentence and annotated sentence under different n-gram, considering TF-IDF (Salton and Buckley 1988) with n-gram. Thus, we expand label fusing from word

level to n-grams-level, where the reliability weight of n-grams TF-IDF starting from $l-$th word sources from averaging ϵ_l of each word in the n-grams, i.e., $\epsilon_l^n = \frac{1}{n}\sum_l^{l+n-1}\epsilon_l$. Thus, each TF-IDF weight in the label representation vector is modified at the n-grams-level:

$$\tilde{g}(y_l)^n = \epsilon_l^n g(y_l)^n + (1-\epsilon_l^n)\, g(\hat{p}_l)^n, \tag{8}$$

where $g(y_l)^n$ is TF-IDF of n-grams ground truth starting from $l-$th word and $g(\hat{p}_l)^n$ is that of n-grams prediction. In like manner, NIC still suggests reducing the gradient of noisy words so as to minimize their negative impact.

4 Experiments

4.1 Datasets

Following traditional image captioning methods (Rennie et al. 2017; Li et al. 2020; Zhang et al. 2021), we validate our method using the MS-COCO dataset (Lin et al. 2014) and the Conceptual Caption dataset (Sharma et al. 2018) collected from the Internet. MS-COCO comprises 123,287 images, each with 5 captions. Using the "Karpaty" splitting approach (Karpathy and Fei-Fei 2017), 5,000 images are allocated for validation, 5,000 for testing, and the remainder for training. To intentionally introduce noise into the MS-COCO annotations, we employ two methods: 1) uniform noise, randomly disrupting captions in the train set by a specified ratio (Hu et al. 2023); 2) asymmetric noise, swapping captions of images with similar category labels during training (Huang et al. 2021), e.g., captions of two images will be swapped when they both map any same object. The noise ratio ranges from 0.2 to 0.8 in increments of 0.2. Conceptual Caption (Sharma et al. 2018) is a large dataset of image-text pairs from the Internet, totaling 3.3 million images, each with a single caption. Due to real internet sources, approximately $3\% \sim 20\%$ of examples include noise. To address computing constraints, we use a subset of Conceptual Captions named CC152K for evaluation, as in (Huang et al. 2021). 1,000 images are allocated for validation, 1,000 for testing, and the remainder for training.

4.2 Implementation Details

NIC is a general framework that applies to almost all existing image captioning methods, where one such state-of-the-art method PureT (Wang, Xu, and Sun 2022) serves as the baseline for our framework. In NIC, we set the model embedding size d_{model} to 512, the number of transformer heads to 8, and the number of refinement encoder and decoder blocks to 3. ADAM (Kingma and Ba 2015) is used as the optimizer, with training epochs set to 25 and 30 for the two stages, and a batch size of 10. The initial learning rate is 5×10^{-6}. The hyperparameters C and τ are initially set to 0.5 and 10 and adaptively adjust once per epoch based on the curvature of the loss curve from the previous epoch. The entire network is trained on an NVIDIA TITAN X GPU. The code is available at https://github.com/njustkmg/NIC.

4.3 Baselines and Evaluation protocol

Compared models are classic state-of-the-art image captioning methods: AoANet (Huang et al. 2019), ORT (Herdade et al. 2019), SGAE (Yang et al. 2019b), M^2 T (Cornia et al. 2020), X-T (Pan et al. 2020), RSTNet (Zhang et al. 2021), DLCT (Luo et al. 2021), PureT (Wang, Xu, and Sun 2022), and SCD-Net (Luo et al. 2023), where AoANet, DLCT, and SCD-Net slightly considered the robustness against noise.

Without loss of generality, all methods are evaluated with metrics such as BLEU (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), ROUGE-L (Lin 2004), CIDEr (Vedantam, Zitnick, and Parikh 2015), and SPICE (Anderson et al. 2016), using the publicly available code. Among them, CIDEr and SPICE are evaluation metrics designed specifically for image captioning tasks.

4.4 Qualitative Analysis

Table 1 presents the results on the uniform noise dataset at an 80% noise ratio. All models are trained in both the cross-entropy and CIDEr optimization stages, with multiple experiments for T-test fairness. The obtained P-values are less than 0.05, signifying a statistically significant difference at a 95% confidence level. The results demonstrate that NIC consistently outperforms all methods in both stages, showing a 6.2/5.3 improvement in CIDEr metrics and a 1.8/1.5 improvement in SPICE metrics compared to the baseline PureT. This underscores NIC's robust ability to generate cross-modal content in high noise ratio scenarios, effectively mitigating the impact of noisy examples. Note that while SCD-Net benefits from using pre-trained CLIP to enhance correlation learning, it suffers from a loss of valuable information in partially matched samples, resulting in inferior performance. In contrast, NIC effectively utilizes both clean and noisy image-text pairs to learn robustly.

To validate the effectiveness of NIC on non-artificially constructed datasets, we conducted experiments on the Conception Caption dataset. Table 2 reports the results of all methods on the real noise dataset CC152K. Similarly, all models learn with training and fine-tuning two stages for fairness. The results show that NIC outperforms the existing state-of-the-art image captioning methods in all metrics, which indicates its excellent robustness in reality.

4.5 Ablation Study

We design several variants of NIC to show the effectiveness of key modules, which respectively remove prediction entropy evaluation (w/o ϵ_1), the cumulative error evaluation (w/o ϵ_2), and the progressive coordination (w/o $h(\gamma)$) module. As shown in the lower part of Table 1: 1) NIC w/o ϵ_1 and w/o ϵ_2 both achieve excellent scores than PureT but lower than NIC, which shows that the entropy of the prediction distribution based on word-region relevance and cumulative error based on contextual information can effectively discriminate noisy words. 2) The performance of using origin reliability weights is worse than that of NIC, which indicates that the progressive coordination strategy ensures a sufficient fitting process on clean samples during the early training and cautious learning in the later stage.

4.6 Influence of the Noise Ratio

To explore the performance of NIC with varying noise ratios, we observe this at four different noise levels: 20%,

Methods	Cross-Entropy Loss								CIDEr Score Optimization							
	B@1	B@2	B@3	B@4	M	R	C	S	B@1	B@2	B@3	B@4	M	R	C	S
AoANet	53.8	37.2	25.5	16.5	16.4	40.2	54.7	11.1	66.9	49.9	34.9	19.3	18.7	46.3	80.1	11.5
ORT	53.4	36.9	25.1	15.2	16.2	39.6	53.7	10.0	65.5	45.7	32.6	18.0	18.2	45.9	79.0	11.0
SGAE	52.6	35.5	24.8	15.0	15.7	39.3	53.2	9.5	65.5	45.7	31.6	17.0	17.2	45.4	78.4	10.3
M^2 T	52.8	36.5	24.9	16.0	16.0	39.1	54.2	10.2	66.1	45.5	32.5	17.0	18.1	45.3	79.1	11.0
X-T	52.8	36.2	25.3	15.1	16.2	39.4	54.6	10.9	66.6	47.2	34.2	18.5	18.6	45.9	79.6	11.2
RSTNet	54.6	37.9	26.0	16.6	16.8	40.8	55.3	11.5	67.6	50.5	35.5	22.7	19.4	47.0	81.0	12.1
DLCT	55.1	38.4	27.5	17.2	17.6	41.3	55.7	11.7	68.2	51.1	36.2	23.5	20.7	47.7	82.4	12.6
PureT	55.4	38.4	27.7	17.5	17.9	41.7	56.2	11.9	68.7	51.7	36.8	23.7	20.8	48.1	83.0	12.8
SCD-Net	55.3	38.5	26.9	16.8	17.3	40.9	55.9	11.0	69.1	51.7	34.8	22.7	20.0	47.4	82.1	12.3
NIC	**58.1**	**41.2**	**29.8**	**19.4**	**20.3**	**46.3**	**62.4**	**13.7**	**71.5**	**55.1**	**38.8**	**26.3**	**23.5**	**50.6**	**88.3**	**14.3**
w/o ϵ_1	57.5	40.6	29.0	18.8	19.6	45.1	60.1	12.8	70.9	54.3	37.8	25.2	22.5	49.6	87.0	13.7
w/o ϵ_2	57.0	39.8	28.4	18.3	18.9	43.9	58.2	12.3	70.0	53.1	37.4	24.7	21.8	48.6	84.7	13.2
w/o $h\,(\gamma)$	57.2	40.4	28.7	18.5	19.5	45.0	58.5	12.4	70.7	53.7	37.5	24.9	21.9	49.2	85.3	13.3

Table 1: Performance of comparison methods on the uniform noise, where B@N, M, R, C and S are short for BLEU@N, METEOR, ROUGE-L, CIDEr and SPICE scores.

Methods	Cross-Entropy Loss								CIDEr Score Optimization							
	B@1	B@2	B@3	B@4	M	R	C	S	B@1	B@2	B@3	B@4	M	R	C	S
AoANet	23.5	15.2	10.6	8.1	11.0	27.6	84.2	18.1	26.1	16.0	10.6	7.5	11.2	28.5	86.3	18.2
ORT	23.1	14.7	10.3	7.7	10.9	27.4	83.1	17.5	25.2	15.4	10.4	7.2	11.0	28.0	83.7	17.6
SGAE	21.7	14.0	9.5	6.2	9.7	26.4	81.1	17.0	24.5	13.8	9.6	6.2	10.1	27.6	82.2	17.1
M^2 T	22.0	14.1	9.7	6.5	10.0	26.4	82.2	17.0	24.9	14.1	9.9	6.8	10.6	27.8	82.9	17.2
X-T	23.1	14.7	10.0	7.4	10.7	27.5	82.3	17.4	25.8	15.3	10.4	7.4	11.0	28.1	84.4	17.8
RSTNet	23.6	15.2	10.8	8.5	11.1	27.9	85.2	18.5	26.7	16.6	10.9	8.9	11.8	29.2	87.5	18.7
DLCT	24.4	15.9	11.3	9.1	12.1	28.4	85.6	18.7	27.4	17.2	11.8	9.7	13.2	29.9	89.2	19.2
PureT	24.7	15.9	11.6	9.7	12.4	28.8	86.1	19.0	27.9	17.3	12.1	10.1	13.3	30.3	89.8	19.4
SCD-Net	24.0	15.0	10.9	9.3	11.3	27.9	85.9	18.4	27.2	17.0	10.9	9.9	12.8	29.4	88.9	18.7
NIC	**27.8**	**18.8**	**14.6**	**10.8**	**15.2**	**32.5**	**93.6**	**20.5**	**31.2**	**19.3**	**14.8**	**11.1**	**15.4**	**32.9**	**95.5**	**20.9**

Table 2: Performance of comparison methods on CC152K, where B@N, M, R, C and S are short for BLEU@N, METEOR, ROUGE-L, CIDEr and SPICE scores.

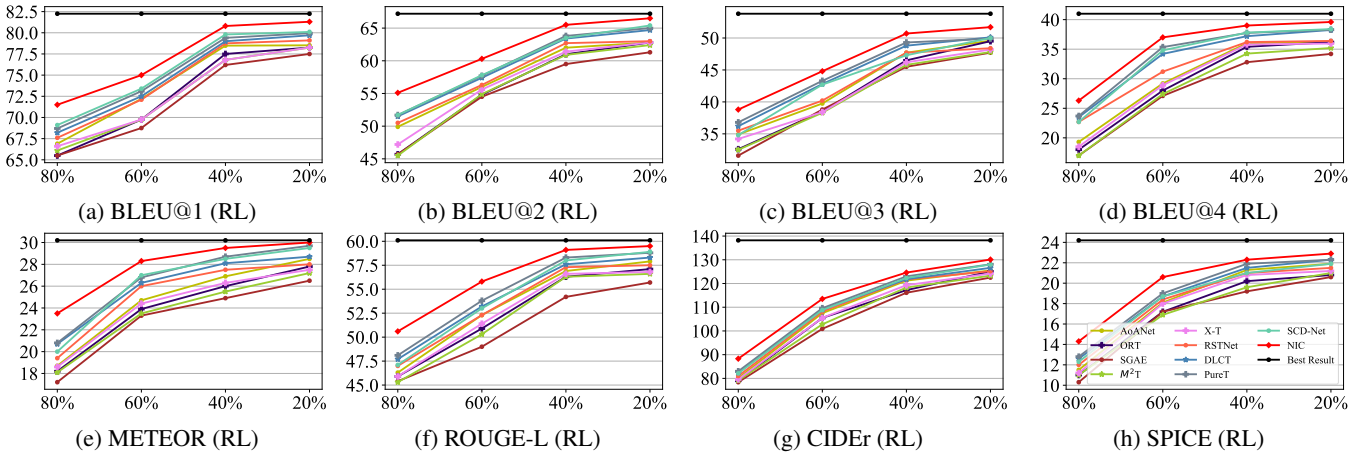
40%, 60%, and 80%. Due to space constraints, only the fine-tuning stage results are shown in Figure 4. NIC outperforms state-of-the-art image captioning methods in all noise ratios. At the same time, the performance of the conventional image captioning method decreases substantially in the high noise ratio scenes, such as 80% and 60% noise ratio, while NIC still presents optimal performance. Therefore, it can be inferred that our method has superior noise immunity. Besides, when compared with the state-of-the-art method (PureT) using clean data (shown with black lines), note that NIC still has competitive performance under 20% noise ratio, which validates again NIC can reasonably utilize the noisy data in a fine-grained manner to optimize the generation of captions.

4.7 NIC Based on Different Captioning Model

Here examines the generality of NIC on different captioning models. Specifically, AoANet, DLCT, and RSTNet act as the base model and are combined with the NIC framework (i.e., AoANet+, RSTNet+, and DLCT+). Among them, AoANet is a traditional encoder-decoder architecture, while RSTNet and DLCT are based on Transformer architecture. Table 3 reports the generalized performance of NIC on the real noisy dataset CC152K. It is evident that our framework leads to performance improvements across all methods, which indicates that NIC can effectively improve the robustness of existing caption models in realistic noise scenarios. Taking DLCT+ as an example, the CIDEr metric shows a 6.5/5.5 improvement, and the SPICE metric improves 1.1/1.4.

4.8 Analysis of Noisy Word Identification

NIC's noise robustness relies on accurately identifying noisy words, a key focus of evaluation. Analyzing 100,000 clean and 100,000 noisy word labels, we visualize the distribution of ϵ_l during training with an 80% noise ratio on the uniform noise dataset (Figure 5 (a)). As training advances, the model's performance improves due to the fine-grained identification and processing of noise, enhancing its ability to recognize noise. In later stages, clean word labels consis-



(a) BLEU@1 (RL) (b) BLEU@2 (RL) (c) BLEU@3 (RL) (d) BLEU@4 (RL)

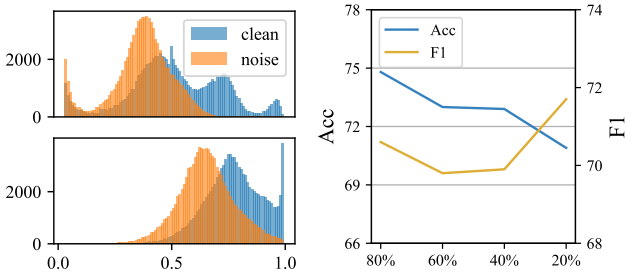
(e) METEOR (RL) (f) ROUGE-L (RL) (g) CIDEr (RL) (h) SPICE (RL)

Figure 4: Captioning performance with different ratios of noisy, where RL represents the results of CIDEr Score Optimization.

Methods	Cross-Entropy Loss							
	B@1	B@2	B@3	B@4	M	R	C	S
AoANet	23.5	15.2	10.8	8.1	11.0	27.6	84.6	18.1
RSTNet	23.9	15.4	10.8	8.8	11.3	27.9	85.2	18.5
DLCT	24.4	15.9	11.3	9.1	12.1	28.4	85.6	18.7
AoANet+	**24.8**	**17.3**	**12.5**	**8.9**	**13.2**	**30.0**	**88.6**	**19.5**
RSTNet+	**25.3**	**17.4**	**12.5**	**9.5**	**13.9**	**30.6**	**89.2**	**19.9**
DLCT+	**26.0**	**17.9**	**13.7**	**10.0**	**14.5**	**32.2**	**92.1**	**19.8**
Methods	CIDEr Score Optimization							
	B@1	B@2	B@3	B@4	M	R	C	S
AoANet	26.1	16.0	10.6	8.5	11.2	28.5	86.9	18.2
RSTNet	26.8	16.6	10.9	8.9	11.9	29.2	87.8	18.7
DLCT	27.4	17.2	11.8	9.7	13.2	29.9	89.2	19.2
AoANet+	**28.9**	**18.0**	**12.9**	**9.6**	**13.5**	**30.7**	**92.2**	**19.9**
RSTNet+	**29.7**	**18.4**	**13.5**	**10.2**	**13.7**	**31.5**	**93.1**	**20.2**
DLCT+	**30.1**	**18.5**	**14.0**	**10.9**	**14.7**	**32.5**	**94.7**	**20.6**

Table 3: Performance of RIC with different caption model on CC152K, where B@N, M, R, C and S are short for BLEU@N, METEOR, ROUGE-L, CIDEr and SPICE.

tently show higher epsilon values than noisy ones, validating the effectiveness of the measurement criterion. Figure 5 (b) illustrates classification accuracy for noisy and clean words based on reliability weight under various noise ratios on the asymmetric noise dataset. While NIC doesn't perform direct binary classification, the model's ability to distinguish them is assessed by manually setting threshold values. The lines consistently show high and smooth performance, affirming the efficacy and robustness of our evaluation criteria. Additionally, as a denoising training framework applicable to various baselines with minimal extra parameters, NIC's complexity primarily arises from entropy and cross-entropy calculations during reliable weight assessment, which depends on the inherent baseline complexity. The overall time complexity is $O(M(M^2 d_{model} + M d_{model}^2 + |Z|O))$, where M is the number of visual regions, $|Z|$ is the vocabulary size.



(a) (b)

Figure 5: Analysis of noisy word identification. (a) illustrates the reliability weight distribution histogram at various training stages, with the upper part depicting the early stage and the lower part representing the mid-to-late stage. (b) reports accuracy and F1 score under varying noise ratios.

5 Conclusion

The noise-label problem has gained widespread attention due to its prevalence. Compared with traditional category labels, sentences containing contextual and semantic information are more complex and challenging, yet research exploring noisy image captioning tasks is insufficient. In this regard, we propose a Noisy-aware Image Captioning (NIC) method that takes into account fine-grained word information instead of explicitly weighting overall examples. In summary, NIC explores mismatched words by adaptively evaluating the reliability weight of each word-label through inter-modal representativeness and intra-modal discriminativeness. Additionally, it progressively fuses model predictions and original word-labels to construct more reliable pseudo-labels, considering learnable word-label weights and model convergence. Practical results demonstrate NIC's effectiveness in handling diverse noisy scenarios, with easy integration into any state-of-the-art image captioning method.

6 Acknowledgments

Here express our gratitude for the support received for this work. This research was partially funded by the National Key RD Program of China (2022YFF0712100), NSFC (62006118, 62276131), Natural Science Foundation of Jiangsu Province of China under Grant (BK20200460), Jiangsu Shuangchuang (Mass Innovation and Entrepreneurship) Talent Program, Young Elite Scientists Sponsorship Program by CAST, and the Fundamental Research Fund for the Central Universities (NO.NJ2022028, No.30922010317).

References

Albawi, S.; Mohammed, T. A.; and Al-Zawi, S. 2017. Understanding of a convolutional neural network. In *ICET*, 1–6.

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In *ECCV*, 382–398. Amsterdam, US.

Baltrusaitis, T.; Ahuja, C.; and Morency, L. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2): 423–443.

Banerjee, S.; and Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *ACL*, 65–72. Michigan, US.

Chen, X.; and Gupta, A. 2015. Webly Supervised Learning of Convolutional Networks. In *ICCV*, 1431–1439. Santiago, Chile.

Cheng, L.; Zhou, X.; Zhao, L.; Li, D.; Shang, H.; Zheng, Y.; Pan, P.; and Xu, Y. 2020. Weakly Supervised Learning with Side Information for Noisy Labeled Images. In *ECCV*, volume 12375, 306–321. Glasgow, UK.

Cho, K.; van Merrienboer, B.; Gülçehre, Ç.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *EMNLP*, 1724–1734. Doha, Qatar.

Cornia, M.; Stefanini, M.; Baraldi, L.; and Cucchiara, R. 2020. Meshed-Memory Transformer for Image Captioning. In *CVPR*, 10575–10584. Washington, US.

Farhadi, A.; Hejrati, S. M. M.; Sadeghi, M. A.; Young, P.; Rashtchian, C.; Hockenmaier, J.; and Forsyth, D. A. 2010. Every Picture Tells a Story: Generating Sentences from Images. In *ECCV*, volume 6314, 15–29. Heraklion, Crete, Greece.

Greff, K.; Srivastava, R. K.; Koutník, J.; Steunebrink, B. R.; and Schmidhuber, J. 2017. LSTM: A Search Space Odyssey. *IEEE Trans. Neural Networks Learn. Syst.*, 28(10): 2222–2232.

Herdade, S.; Kappeler, A.; Boakye, K.; and Soares, J. 2019. Image Captioning: Transforming Objects into Words. In *NeurIPS*, 11135–11145. British Columbia, UK.

Hossain, M. Z.; Sohel, F.; Shiratuddin, M. F.; and Laga, H. 2019. A Comprehensive Survey of Deep Learning for Image Captioning. *ACM Comput. Surv.*, 51(6): 118:1–118:36.

Hu, P.; Huang, Z.; Peng, D.; Wang, X.; and Peng, X. 2023. Cross-Modal Retrieval With Partially Mismatched Pairs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(8): 9595–9610.

Huang, L.; Wang, W.; Chen, J.; and Wei, X. 2019. Attention on Attention for Image Captioning. In *ICCV*, 4633–4642. Seoul, Korea.

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 *NeurIPS*, 29406–29419. virtual.

Kang, W.; Mun, J.; Lee, S.; and Roh, B. 2023. Noise-aware learning from web-crawled image-text data for image captioning. In *ICCV*, 2942–2952.

Karpathy, A.; and Fei-Fei, L. 2017. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4): 664–676.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*. San Diego, CA, USA.

Li, J.; Selvaraju, R. R.; Gotmare, A.; Joty, S. R.; Xiong, C.; and Hoi, S. C. 2021. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *NeurIPS*, 9694–9705. virtual.

Li, S.; Kulkarni, G.; Berg, T. L.; Berg, A. C.; and Choi, Y. 2011. Composing Simple Image Descriptions using Web-scale N-grams. In *CoNLL*, 220–228. Portland, Oregon, USA.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; Choi, Y.; and Gao, J. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *ECCV*, volume 12375, 121–137. Glasgow, UK.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*, volume 8693, 740–755. Zurich, Switzerland.

Luo, J.; Li, Y.; Pan, Y.; Yao, T.; Feng, J.; Chao, H.; and Mei, T. 2023. Semantic-conditional diffusion networks for image captioning. In *CVPR*, 23359–23368.

Luo, Y.; Ji, J.; Sun, X.; Cao, L.; Wu, Y.; Huang, F.; Lin, C.; and Ji, R. 2021. Dual-level Collaborative Transformer for Image Captioning. In *AAAI*, 2286–2293. virtual.

Lyu, Y.; and Tsang, I. W. 2020. Curriculum Loss: Robust Learning and Generalization against Label Corruption. In *ICLR*. Addis Ababa, Ethiopia.

Pan, Y.; Yao, T.; Li, Y.; and Mei, T. 2020. X-Linear Attention Networks for Image Captioning. In *CVPR*, 10968–10977. Seattle, WA, USA.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *ACL*, 311–318. Philadelphia, US.

Qin, Y.; Peng, D.; Peng, X.; Wang, X.; and Hu, P. 2022. Deep Evidential Learning with Noisy Correspondence for Cross-modal Retrieval. In *MM*, 4948–4956. Lisboa, Portugal.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, volume 139, 8748–8763. virtual.

Reed, S. E.; Lee, H.; Anguelov, D.; Szegedy, C.; Erhan, D.; and Rabinovich, A. 2015. Training Deep Neural Networks on Noisy Labels with Bootstrapping. In *ICLR*. San Diego, CA, USA.

Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-Critical Sequence Training for Image Captioning. In *CVPR*, 1179–1195. Honolulu, HI.

Salton, G.; and Buckley, C. 1988. Term-Weighting Approaches in Automatic Text Retrieval. *Inf. Process. Manag.*, 24(5): 513–523.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *ACL*, 2556–2565. Melbourne, Australia.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *NeurIPS*, 5998–6008. California, US.

Vedantam, R.; Zitnick, C. L.; and Parikh, D. 2015. CIDEr: Consensus-based image description evaluation. In *CVPR*, 4566–4575. Massachusetts, US.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *CVPR*, 3156–3164. Massachusetts, US.

Wang, X.; Hua, Y.; Kodirov, E.; Clifton, D. A.; and Robertson, N. M. 2021. ProSelfLC: Progressive Self Label Correction for Training Robust Deep Neural Networks. In *CVPR*, 752–761. virtual.

Wang, Y.; Ma, X.; Chen, Z.; Luo, Y.; Yi, J.; and Bailey, J. 2019. Symmetric Cross Entropy for Robust Learning With Noisy Labels. In *ICCV*, 322–330. Seoul, Korea (South).

Wang, Y.; Xu, J.; and Sun, Y. 2022. End-to-End Transformer Based Model for Image Captioning. In *AAAI*, 2585–2594. virtual.

Yang, M.; Zhao, W.; Xu, W.; Feng, Y.; Zhao, Z.; Chen, X.; and Lei, K. 2019a. Multitask Learning for Cross-Domain Image Captioning. *IEEE Trans. Multim.*, 21(4): 1047–1061.

Yang, X.; Tang, K.; Zhang, H.; and Cai, J. 2019b. Auto-Encoding Scene Graphs for Image Captioning. In *CVPR*, 10685–10694. California, US.

Yang, Y.; Sun, Z.; Zhu, H.; Fu, Y.; Zhou, Y.; Xiong, H.; and Yang, J. 2023. Learning Adaptive Embedding Considering Incremental Class. *IEEE Trans. Knowl. Data Eng.*, 35(3): 2736–2749.

Yang, Y.; Wei, H.; Zhu, H.; Yu, D.; Xiong, X.; and Yang, J. 2022. Exploiting Cross-Modal Prediction and Relation Consistency for Semisupervised Image Captioning. *IEEE Transactions on Cybernetics*.

Yang, Y.; Wu, Y.-F.; Zhan, D.-C.; Liu, Z.-B.; and Jiang, Y. 2019c. Deep robust unsupervised multi-modal network. In *AAAI*, volume 33, 5652–5659.

Yang, Y.; Zhan, D.; Jiang, Y.; and Xiong, H. 2020. Reliable multi-Modal learning: a survey. *Journal of Software*, 32(4): 1067–1081.

Yang, Y.; Zhan, D.; Wu, Y.; Liu, Z.; Xiong, H.; and Jiang, Y. 2021. Semi-Supervised Multi-Modal Clustering and Classification with Incomplete Modalities. *IEEE Trans. Knowl. Data Eng.*, 33(2): 682–695.

You, Q.; Jin, H.; Wang, Z.; Fang, C.; and Luo, J. 2016. Image Captioning with Semantic Attention. In *CVPR*, 4651–4659. Las Vegas, NV, USA.

Zhang, X.; Sun, X.; Luo, Y.; Ji, J.; Zhou, Y.; Wu, Y.; Huang, F.; and Ji, R. 2021. RSTNet: Captioning With Adaptive Attention on Visual and Non-Visual Words. In *CVPR*, 15465–15474. virtual.

Zhang, Z.; Wu, Q.; Wang, Y.; and Chen, F. 2019. High-Quality Image Captioning With Fine-Grained and Semantic-Guided Visual Attention. *IEEE Trans. Multim.*, 21(7): 1681–1693.