Efficient Modeling of Future Context for Image Captioning

Zhengcong Fei, Junshi Huang*, Xiaoming Wei, Xiaolin Wei

Meituan

Beijing, China

{feizhengcong,huangjunshi,weixiaoming,weixiaolin}@meituan.com

ABSTRACT

Existing approaches to image captioning usually generate the sentence word-by-word from left to right, with the constraint of conditioned on local context including the given image and history generated words. There have been many studies target to make use of global information during decoding, e.g., iterative refinement. However, it is still under-explored how to effectively and efficiently incorporate the future context. To respond to this issue, inspired by that Non-Autoregressive Image Captioning (NAIC) can leverage two-side relation with modified mask operation, we aim to graft this advance to the conventional Autoregressive Image Captioning (AIC) model while maintaining the inference efficiency without extra time cost. Specifically, AIC and NAIC models are first trained combined with shared visual encoders, forcing the visual encoder to contain sufficient and valid future context; then the AIC model is encouraged to capture the causal dynamics of cross-layer interchanging from NAIC model on its unconfident words, which follows a teacher-student paradigm and optimized with the distribution calibration training objective. Empirical evidences demonstrate that our proposed approach clearly surpass the state-of-the-art baselines in both automatic metrics and human evaluations on the MS COCO benchmark. The source code is available at: https://github.com/feizc/Future-Caption.

CCS CONCEPTS

• **Computing methodologies → Natural language generation**; *Learning latent representations*; Computer vision tasks.

KEYWORDS

Future Context; Image Captioning; Non-autoregressive Decoding; Causal Dynamics Calibration
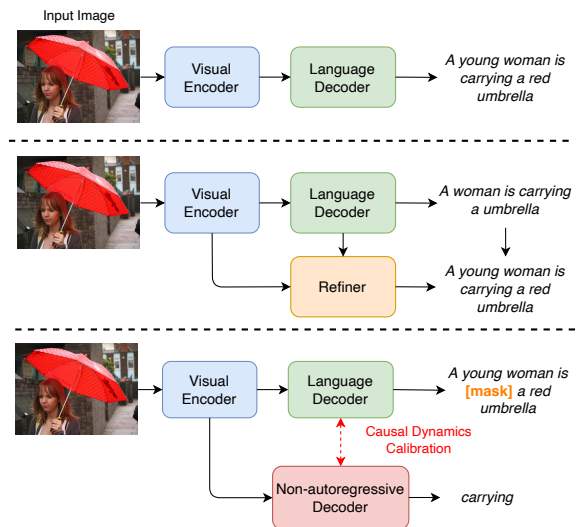
*Corresponding author.

Figure 1: Overview of conventional image captioning, refinement-based image captioning, and our future context modeling with causal dynamics calibration from non-autoregressive decoder. Note that the non-autoregressive decoder is not involved at the inference stage to maintain computation efficiency.

1 INTRODUCTION

Image captioning, which aims to describe the image content with natural language, has seen rapid development in the past several years [6]. In a conventional image captioning system, an visual encoder first transform the given image into a sequence of intermediate hidden representations, based on which, a language decoder generate the sentence word by word. Such encoder-decoder paradigm is usually implemented by CNN-LSTM [53, 58] or Transformer [51] network architecture, and optimized with teacher forcing objectives [3, 9, 22, 24, 40, 63]. Despite its success, the autoregressive structure of the left-to-right manner makes the model only access to the local context, *i.e.*, the previously generated words as well as the given image, at each decoding step. Such a unidirectional property makes the models unable to exploit global context effectively, yielding an unsatisfied description. [54, 55, 67].

To address this issue, many researchers have attempted to exploit the global information during sentence generation. Typically, refinement-based method is introduced [29, 47–49, 56, 57, 59, 62], which typically consists of two networks: the first network is usually a primary generator or an image-text retrieval, which is used to generate or retrieve a coarse related template; A refiner is in series to generate the final caption by attending to the sentence

produced before. Such an iterative refinement operation can help the model to look at both past and future semantic context and thus improve decoding at every time step. However, most of the works rely on multi-pass decoding or specially customized decoding algorithms which leads to a significant increase in training and inference costs. On the other hand, modeling the global context in the reverse direction by pairing the conventional left-to-right image captioning model with a right-to-left auxiliary model is also delivered [11, 48, 50, 55]. However, in these methods, the modeling of reverse context is still conditioned on the local context with a separate network and they cannot sufficiently encourage the image captioning model to exploit a truly flexible global context.

In pursuit of effectively and efficiently incorporating global information into image captioning models, we conduct well-designed pilot experiments, and find some interesting phenomenons: i) Even conditioned on absolutely correct context, i.e., historical words and given image, there is a certain proportion of ground-truth words that the image captioning model predicts with relatively low probabilities. ii) The probability assigned to the ground-truth words from image captioning models differs with the captioning length. In contrast, with the factorized probability modeling, a good image captioning model should endow the highest probability to correct words according to accurate historical information. Consistent with [19, 66, 67], we believe that the reasonable cause of this phenomenon is that the image captioning model cannot confidently predict these words according to only the local context. Therefore, it should be improved for the image captioning model on these unconfident words with sufficient distribution calibration.

In this paper, we introduce efficient modeling of future context information for image captioning, referred to as FutureCap. In general, the architecture of the original autoregressive image captioning (AIC) model is kept untouched and jointly optimized with an additional mask-based non-autoregressive image captioning (NAIC) model [17, 18, 20], which is essentially cross-modal understanding and contains the global context. As shown in Figure 1, the AIC model and NAIC model are first trained combined in multi-task learning with sharing the visual encoders. The visual encoder is additionally supervised by the signal from the NAIC decoder to include sufficient future context information; Then, we employ casual dynamics calibration that pushes the student AIC model faithfully learn the causal effect of teacher NAIC model representations on unconfident output, with cross-layer interchanging aligning. This further help the AIC model leverage knowledge to information dynamics. Experimentally, we evaluate our approach to the MS COCO dataset. According to both automatic metrics and human evaluations, the captioning models equipped with future context modeling evidently outperform baselines. The major **contributions** of our paper are as follows:

- We focus on the efficient modeling of future information for better image caption decoding and analyze the necessity of global context clearly with pilot experiments.
- We introduce causal dynamics calibration that encourages the student AIC model to learn the interchange aligning from teacher NAIC model on unconfident words and adjust knowledge routing with share visual encoder, to more effectively exploit the future contextual information.

- Experiments on the MS COCO dataset demonstrate that image captioning models equipped with our future context modeling framework significantly outperforms the one without it. More encouragingly, as most of the previous literature improves the performance by increasing the model capacity, our approach represents a new optimization paradigm that leads to no additional inference cost.

2 BACKGROUND AND PILOT ANALYSIS

To investigate the potential impact of future context in image captioning, we first describe the basic architectures of conventional autoregressive and non-autoregressive image caption models, both of which follows on Transformer-based encoder-decoder paradigm. After that, we conduct pilot experiments as well as empirical analyses on the effects of context information for caption decoding.

2.1 Model Architecture

Generally, AIC and NAIC models hold the same visual encoder architecture while differing in decoders for their mask matrix of self-attention mechanisms and prediction manners.

Visual Encoder. The visual encoder aims to learn the high-level visual representations of the given image, which includes L same network layers. Each layer consists of two sub-layers: a self-attention sub-layer and a position-wise feed-forward network sub-layer. The input of the network layer is the hidden states of the previous layer, on which the multi-head scaled dot-product attention computation is performed. Assuming that h_e^l presents the hidden states of the l-th encoder layer, the visual encoder layer can be computed as:

$$s^l = \text{SelfAttention}(h_e^{l-1}, h_e^{l-1}, h_e^{l-1}), \tag{1}$$

$$h_e^l = \text{FeedForward}(s^l). \tag{2}$$

Layer normalization with residual connection is added after both two sub-layers. Note that h_e^0 is initialized as the patch embedding of the extracted image region features, and the hidden states of the L-th layer h_e^L are served as input to the language decoder.

Language Decoder. The decoder of the AIC and NAIC models are introduced separately. For the autoregressive decoder, which usually consists of three sub-layers: a masked self-attention sub-layer, a cross-attention sub-layer, and an FFN sub-layer. In particular, to maintain the autoregressive generated property at each time step, the masked self-attention sub-layer performs self-attention with a causal attention mask to prevent the decoder from seeing subsequent words. To generate the hidden states h_d^l of the l-th decoder layer, the autoregressive decoder can be formulated as:

$$s^l = \text{MaskSelfAttention}(h_d^{l-1}, h_d^{l-1}, h_d^{l-1}), \tag{3}$$

$$c^l = \text{CrossAttention}(s^l, h_e^L, h_e^L), \tag{4}$$

$$h_d^l = \text{FeedForward}(c^l). \tag{5}$$

Layer normalization with residual connection is also added after each sub-layers. Finally, with the given image x, the generated sentence $w_{<t}$ and the learned top-layer hidden states $h_{d,t}$, the decoder models the probability distribution as:

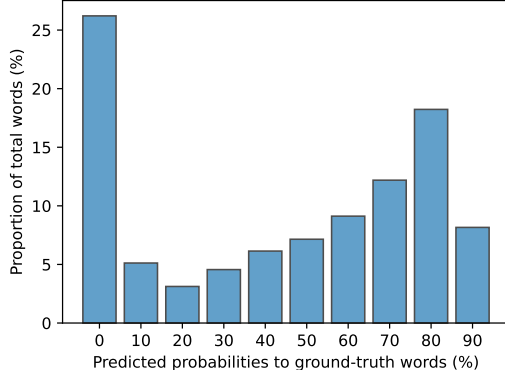$$p_{AIC}(w_t|w_{<t}, x) = \text{Softmax}(Wh_{d,t}), \tag{6}$$

Figure 2: Predicted probability of the different ground-truth words on the training set of MS COCO dataset.
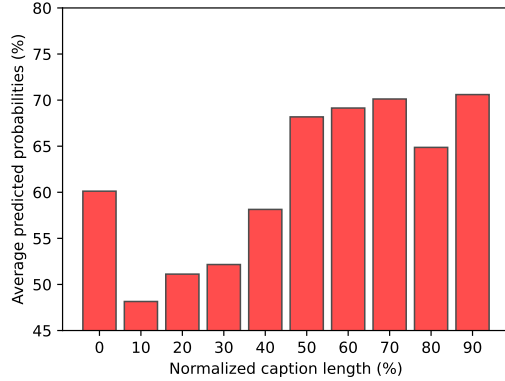

Figure 3: Average predicted probability of the ground-truth words for the different normalized sentence length on the training set of MS COCO dataset.

where W denotes the learnable parameter.

For the non-autoregressive decoder, which aims to predict a set of masked target words S_m given an image x and a set of observable target words S_o. The NAIC decoder also contains the same L identical layers, each of which also includes a self-attention sub-layer, a cross-attention sub-layer, and a feedforward sublayer. Unlike the masked self-attention sub-layer of the AIC decoder, the attention mask is removed in the NAIC decoder. Finally, with the learned top-layer hidden states $\tilde{h}_d$ of the NAIC decoder and partially observed sentence S_o, the predicted probability distribution for every masked word $w_t \in S_m$ can be calculated as:

$$p_{NAIC}(w_t|S_o, x) = \text{Softmax}(\tilde{W}\tilde{h}_{d,t}), \tag{7}$$

where $\tilde{W}$ is the learnable parameter. Note that since the decoder of the NAIC model takes S_o rather than $w_{<t}$ as input, which includes both history and future words with respect to every masked target word, it should embody the global contextual information.

2.2 Are History Contexts Enough for Prediction?

A high-quality image captioning model is supposed to endow the highest probabilities of the ground-truth words based on correct historical context. Delicate experiments are conducted in this section to explore the disadvantage of conventional transformer image captioning with a teacher-forcing training framework.

Experimental Setting. Experimentally, we adopt the basic configuration of the Transformer-based image captioning model without the mesh-memory module, which is publicly available at GitHub [1]. To be specific, the model is comprised of 6 standard transformer layers of visual encoder and language decoder. Moreover, the regional features of images extracted from faster r-cnn [44] on the backbone of ResNet [21] are utilized to retrain the image captioning model with the current configuration [51]. For the training process, the image captioning model is first trained with cross-entropy loss and then fine-tuned with sentence-level self-critical reward [46] follow the default training settings with Adam [30] optimizer on the MS COCO [6] dataset of Karpathy training split.

After obtaining a fully-trained image captioning model, we record the predicted probability of the ground-truth words given the correct context including image and the previous subsentence in the MS COCO training set. In order to characterize the results, we plot the proportion of total words in different predicted probability chunks in Figure 2. Meantime, we also plot the average predicted probability of the corresponding ground-truth words for different caption lengths in Figure 3. Here we normalized the caption to eliminate the influence of absolute sentence length.

Results Discussion. According to the estimation results of Figure 2, we can find that even provided with the totally correct context, there is an obvious portion of ground-truth words that the conventional image captioning model predicts with relatively low probabilities. For instance, the model predicts 25.67% ground-truth words with probabilities chunk between 0.0-0.1. The reasonable cause of this phenomenon is that **image captioning model cannot confidently predict these ground-truth words according to only the local context of image and history words**.

To further prove where the low confidence words locate, we calculate the average predicted probability for various locations in the total length. We can see that with the relative length increase, *i.e.*, the generation position moves from left to right, and the average probability of ground-truth words increases gradually. We attribute that with the generated sentence length increase, the determined context increase, and future context becomes less, which results in the strengthening of model confidence. According to the experimental results, it is natural to consider improving the image captioning model on these correct while unconfident words with effective future information to assist the current decision.

3 METHODOLOGY

Based on the above analysis, we believe that modeling future information in image captioning, especially for unconfident words, is necessary. To this end, we propose a new framework for efficient modeling of future context, named FutureCap, that tries to employ

---

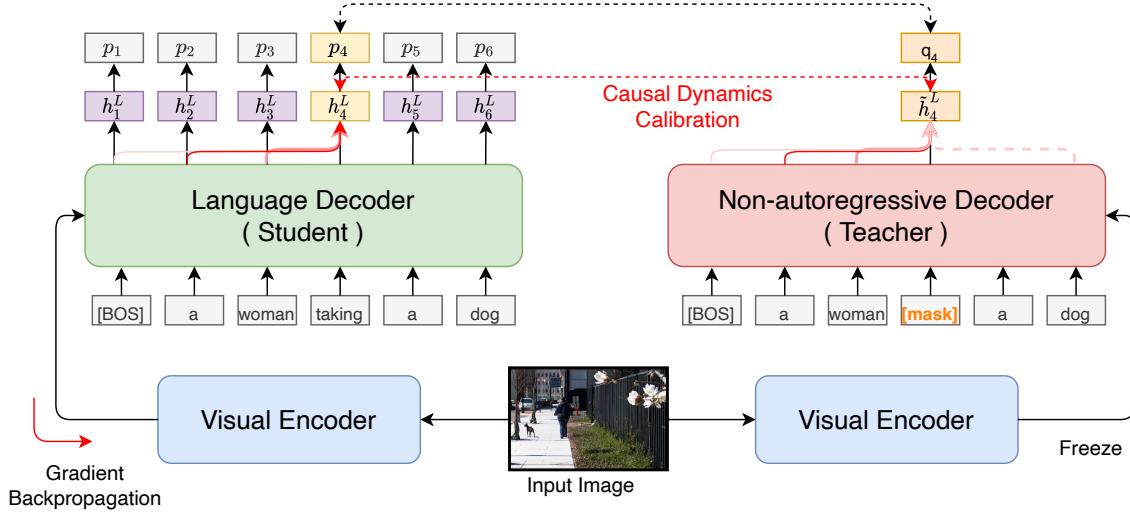[1]https://github.com/aimagelab/meshed-memory-transformer

Figure 4: Illustration of future context modeling with non-autoregressive decoder for language decoder. Assuming that the confidence of predicted word $\{p_4\}$ from language decoder is lower than threshold ϵ, the masked words set S_m becomes $\{w_4\}$. Thus the input S_o to the NAIC decoder is $\{w_1, w_2, w_3, [mask], w_5, w_6\}$, and the output hidden states $\tilde{h}_4^L$ and probability q_4 is used to calibrate the causal dynamics of original p_4. Note that the total parameters of the NAIC model is freezed.

a mask-based non-autoregressive image captioning decoder to enhance the conventional image captioning model according to its prediction confidence. Overall, we first train the AIC model and NAIC model with a shared visual encoder to acquire an image encoding supervision. Then, we employ the NAIC model as a teacher to improve the unconfident word of the student AIC model through causal dynamics calibration.

3.1 Shared Visual Encoder Supervision

To encourage the visual encoder containing sufficient global information, we first train the image captioning model and NAIC model with shared visual encoder in a multi-task manner and optimize the combined training objective as follows:

$$L_{VE}(\theta_{VE}, \theta_{AIC}, \theta_{NAIC}) = \lambda L_{AIC} + (1 - \lambda)L_{NAIC}, \quad (8)$$

where θ_{VE}, θ_{AIC} and θ_{NAIC} denote the parameters of the shared visual encoder, the autoregressive language decoder, and the mask-based non-autoregressive decoder, respectively. λ is a balancing factor between two losses. As the visual encoder is additionally supervised by the signal from the mask-based NAIC decoder, the AIC model is able to disentangle the future information from the extracted visual representation. In between, the AIC model is first optimized through the time-wise cross-entropy loss:

$$L_{AIC}(\theta_{VE}, \theta_{AIC}) = -\sum_{t=1}^{|S|} \log p(w_t|w_{<t}, I). \quad (9)$$

And then fine-tuning using with CIDEr score reward r and mean baseline b. The gradient expression for SCST [46] training is,

$$\nabla_{(\theta_{VE}, \theta_{AIC})}L_{AIC} = -((r(w^i) - b) \nabla_{(\theta_{VE}, \theta_{AIC})}\log p_{AIC}(w^i)). \quad (10)$$

For NAIC decoder, we adopt the strategy in [18]. Concretely, we randomly select n words, and replace each selected word with a

special symbol $[mask]$, splitting sentence S into observed set S_o and masked set S_m. We eventually minimize the following training objective for each word in masked sentence S_m as:

$$L_{NAIC}(\theta_{VE}, \theta_{NAIC}) = -\sum_{w_t \in S_m} \log p(w_t|S_o, I). \quad (11)$$

3.2 Causal Dynamics Calibration

We then introduce to use NAIC model as a teacher to transfer the knowledge for the student NAIC model on unconfident words decision, i.e., help the NAIC model to capture and consider more global information from visual representation and generated words. The parameters of the teacher NAIC model are frozen in this stage. Figure 4 depicts the training procedure of this stage with an easy understanding example. Formally, given the image I and the generated ground-truth words $w_{<t}$ at each time step t, we first ask the AIC model make predictions for every word using Equation 6, generating the prior word-level probability distributions $\{p_1, p_2, \ldots, p_{|S|}\}$ and $|S|$ is the sentence length. Then, the masked word set S_m is built, where the predicted probabilities p_t to the corresponding ground-truth words are lower than a threshold value ϵ as: $S_m = \{w_t|p_t \leq \epsilon, 1 \leq t \leq |S|\}$. Next, we obtain the observed set S_o for NAIC model input by replacing those selected low-confident ground-truth words in the original sentence S with a special symbol $[mask]$. Note that the equation $S = S_o \cup S_m$ is always true in our framework. Next, we can obtain the predicted probability distribution $\hat{q}_t$ from the teacher NAIC model for every word in S_m using Equation 7.

Once get the knowledge routing of the AIC and NAIC model, to improve the decision on the set S_m of its unconfident words, we introduce *causal dynamics calibration* to assist the future context

---

Algorithm 1: Causal Dynamics Calibration Algorithm between student AIC model and teacher NAIC model

---

Input: Unconfident masked data S_m, student AIC model p_{AIC} with output neural N, teacher NAIC model p_{NAIC}, neural alignment g

1 Fix the parameters of p_{NAIC};

2 **while** *not converged* **do**

3   **for** w_t *in* S_m **do**

4   N_S = Sample_student_neurons(N);

5       $N_T = g(N_s)$;

6   Compute the causal dynamic difference

7       $\| \text{GET}(p_{AIC}, N_S, w_t) - \text{GET}(p_{NAIC}, N_T, w_t) \|_2^2$;

8   Compute KD loss $\text{KL}(q_t \| p_t)$;

9   Compute combined loss L_{CDC} ;

10   Loss backward;

11   Step optimizer;

12   **end**

13 **end**

---

modeling of AIC model. The detailed progress is shown in Algorithm 1. In between, GET operation is defined as an activation-value retriever for a neural model. Given a model p contain a set of neurons N, i.e., internal representations, and input context c_t including image x and generated word $w_{<t}$, $\text{GET}(p, N, c_t)$ is the set of weight values that neural N takes on when processing the context c_t. For context c_t, N_s is the set of neuros from student AIC model, we can get the interchanging alignment loss as:

$$L_{IA}(\theta_{VE}, \theta_{AIC}) = \sum_{w_t \in S_m} \|\text{GET}(p_{AIC}, N, w_t) \\ - \text{GET}(p_{NAIC}, g(N), w_t)\|_2^2 \tag{12}$$

where $g(\cdot)$ is the mapping function for sampled neurons except the future neurons in teacher NAIC model. Similar to the conventional knowledge distillation [23], we also restricted the output distribution for unconfident words with KL-divergence as:

$$L_{KL}(\theta_{VE}, \theta_{AIC}) = \sum_{w_t \in S_m} \text{KL}(q_t \| p_t), \tag{13}$$

The final training objective for the student AIC model is a combination of the three terms reviewed above as:

$$L_{CDC}(\theta_{VE}, \theta_{AIC}) = L_{KL}(\theta_{VE}, \theta_{AIC}) + L_{IA}(\theta_{VE}, \theta_{AIC}) \\ - \sum_{w_t \in S_o} \log p_t, \tag{14}$$

where the last term makes the AIC model stable to the high confidence ground-truth words. By doing so, we can fully strengthen the ability of the AIC model to leverage the global context contained in the NAIC. On the other hand, to avoid making the student model rely heavily on the decision of the teacher NAIC model, we also employ a teacher annealing strategy to linearly decrease the knowledge distillation to ground-truth supervision of sentence-level reward [46] throughout training. Note that the NAIC model is not involved at the inference stage to keep efficiency.

Table 1: Performance comparisons of our FutureCap model and other state-of-the-art image captioning models with different evaluation metrics on the MS COCO Karpathy test set. All values are reported as a percentage (%).

	B-1	B-4	M	R	C	S
LSTM-A [61]	78.6	35.5	27.3	56.8	118.3	20.8
Up-Down [3]	79.8	36.3	27.7	56.9	120.1	21.4
GCN-LSTM [60]	80.5	38.2	28.5	58.3	127.6	22.0
AoANet [24]	80.2	38.9	29.2	58.8	129.8	22.4
$\mathcal{M}^2$ Transformer [9]	80.8	39.1	29.2	58.6	131.2	22.6
X-LAN [40]	80.8	39.5	29.5	59.2	132.0	23.4
DPA [35]	80.3	40.5	29.6	59.2	133.4	23.3
GET [26]	81.5	39.5	29.3	58.9	131.6	22.8
DLCT [37]	81.4	39.8	29.5	59.1	133.8	23.0
RSTNet [63]	81.8	40.1	29.8	59.5	135.6	23.3
FutureCap	**82.2**	**40.3**	**30.1**	**59.8**	**136.3**	**23.8**

4 EXPERIMENTS

4.1 Experimental Preparation

Dataset. We evaluate the proposed method on the MS COCO [6], which is a standard estimation benchmark for image captioning tasks. To be consistent with previous work [9, 24], we adopted the Karpathy split [28] that contains 113,287 training images and 5,000 images for validation and test splits, respectively. Each image corresponds with 5 different captions. We omit words that occur less than 5 times and the vocabulary size is 10,369 words. Image features are extracted with CLIP [3] for 512-dim vectors.

Evaluation Metrics. Following the common paradigm, we utilize five metrics to comprehensively estimate the captioning performance: BLEU-N [41], METEOR [31], ROUGE [34], CIDEr [52], and SPICE [2]. We denote as B-N, M, R, C, and S for simplicity.

Implementation Details. Our implementation is based on Pytorch and repository [9] to build the model under Transformer-base configuration with the memory module, where AIC and NAIC hold the identical architecture. Concretely, both the network is comprised of 6 visual encoder and 6 language decoder layers, each with 512 as hidden size, the FFN sublayers of 2,048 dimensions, and 8 heads in multi-head attentions. We set the dropout rate to 0.1. The neurons are sampled from the top-layer decoder with a unifying distribution. For parameter updating, we employ the Adam optimizer [30] with the default setting. As for learning rate schedule, we adopt the same strategy as [9, 51] and set warm-up steps to 4,000. In the first stage, we train all models by sharing their encoders for 300k steps. In the second stage, we separate their encoders and fix the parameter of the NAIC model. Then, the AIC model is solely optimized with the fixed NAIC by additional 200k steps.

4.2 Comparison with State-of-the-Art Models

Performance on MS COCO. We compare the results of our FutureCap model with those of several recent image captioning models trained without large-scale vision-and-language pre-training on the offline MS COCO dataset. The evaluation results are listed in Table 1. First, we can see that FutureCap surpasses the original

Table 2: Leaderboard of different image captioning models on the online MS COCO test server.

	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE		CIDEr	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
Up-Down [3]	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
AoANet [24]	81.0	95.0	65.8	89.6	51.4	81.3	39.4	71.2	29.1	38.5	58.9	74.5	126.9	129.6
$\mathcal{M}^2$ Transformer [9]	81.6	96.0	66.4	90.8	51.8	82.7	39.7	72.8	29.4	39.0	59.2	74.8	129.3	132.1
X-Transformer [40]	81.9	95.7	66.9	90.5	52.4	82.5	40.3	72.4	29.6	39.2	59.5	75.0	131.1	133.5
GET [26]	81.6	96.1	66.5	90.9	51.9	82.8	39.7	72.9	29.4	38.8	59.1	74.4	130.3	132.5
DPA [35]	81.8	96.3	66.5	91.2	51.9	83.2	39.8	73.3	29.6	39.3	59.4	75.1	130.4	133.7
RSTNet [63]	82.1	96.4	67.0	91.3	52.2	83.0	40.0	73.1	29.6	39.1	59.5	74.6	131.9	134.0
FutureCap	**82.4**	**96.7**	**67.3**	**91.8**	**52.6**	**83.8**	**40.3**	**74.0**	**29.6**	**39.2**	**59.6**	**74.9**	**132.9**	**135.3**

Table 3: Ablation studies on the MS COCO test set.

Model	B-1	B-4	M	R	C	S
FutureCap	**82.2**	**40.3**	**30.1**	**59.8**	**136.3**	**23.8**
w/o. VES	81.6	40.0	29.5	59.5	135.2	23.6
w/o. CDC	81.4	39.9	29.4	59.4	134.8	23.5
w/. KL	82.1	40.1	29.9	59.6	135.9	23.6

memory-incorporated Transformer by +0.6 BLEU-4 and 7.1 CIDEr scores, respectively, verifying that modeling the future information brings a significant performance improvement. Next, it is encouraging that our proposed framework outperforms the most recent competitive models. As it can be observed, our proposal reaches 136.3 CIDEr points, beating almost all the compared approaches. It is encouraging that our strategy can be combined with other advance improved strategies as well as untouched the internal structures. What's more, as most of the previous literature has boosted caption quality by increasing the model capacity, which leads to an extra burden for application devices, our approach represents an outliner in this trend and demonstrates that state-of-the-art CIDEr levels can be obtained even with a very lightweight efficient model.

Online evaluation. We also report the performance of our method on the online MS COCO test server. In this case, we employ an ensemble of four models trained with the same configuration of NAIC decoder assistance, for which ground-truth annotations are not publicly available. Comparison results with the top-performing approaches of the leaderboard are reported in Table 2. As it can be seen, our method surpasses the current state-of-the-art model on all metrics, achieving an advancement of 1.3 CIDEr points with respect to the best performer.

4.3 Model Analysis

Ablation Study. To better understand the influence of each design in our FutureCap model, we conduct ablation studies on the offline MS COCO dataset. Table 3 reports the evaluation results on the testing set. We first validate the necessity of shared visual encoder supervision by training the AIC and NAIC model using the separate visual encoder, respectively, denoted as "w/o. VES". The automatic metrics of final performance decrease by a 1.1 CIDEr score. This shows that the visual encoder of the AIC model benefits
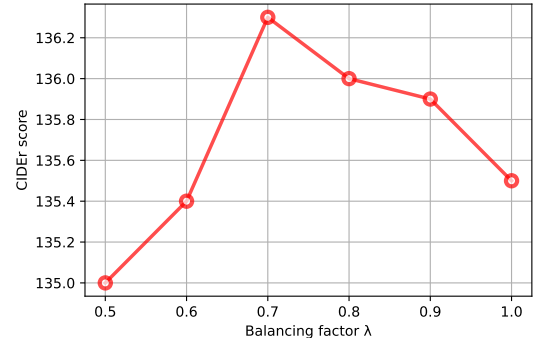
Figure 5: The evaluated CIDEr scores according to the combine training stage on MS COCO offline test set with different λ, the balancing factor in different loss.

a lot from the global supervision information through joint training with the NAIC decoder. As for "w/o. CDC", which means not performing causal dynamics calibration on any target words at the fine-tuning stage, its performance also decreases, e.g. 0.4 for the BLEU-4 score and 1.5 for the CIDEr score. Moreover, to illustrate the superiority of CDC, we replaced it with the conventional knowledge distillation, i.e., remove the L_{IA} in Equation 14, the results show poor performance. These demonstrate the effect of each part in our design as well as incorporating specific future context into the image captioning model on its unconfident words.

Effect of Hyper-parameters λ and ϵ. There exist important hyper-parameters in the FutureCap framework that we need to tune on the validation set to achieve a good performance, i.e., the balancing factor λ in Equation 8 and the confidence threshold ϵ for determining mask words set S_m. To balance the training of the AIC model and NAIC model at the pre-training stage, we try to select the optimal λ that can bring steady improvements to the AIC model. Specifically, we gradually vary λ from 0.5 to 1.0 with an increment of 0.1 and evaluate the performance on the validation set, the evaluation results are as shown in Figure 5. We can see that the final image captioning model achieves its peak when $\lambda = 0.7$. Hence, λ is set to 0.7 by default. Given the selected λ, at the fine-tuning calibration stage, we also analyze the impact of ϵ on the validation
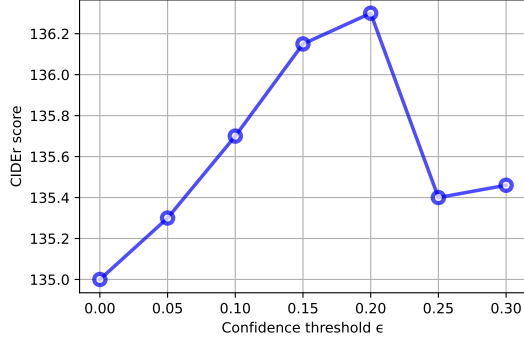
Figure 6: The evaluated CIDEr scores on the MS COCO offline test set with different value of ϵ, the confidence threshold for masked words.

Table 4: Performance comparisons of incorporating different distribution calibration strategies on the MS COCO test set.

Model	B-1	B-4	M	R	C	S
FutureCap	**82.2**	**40.3**	**30.1**	**59.8**	**136.3**	**23.8**
Random	81.9	40.2	29.7	59.6	135.8	23.7
Highest	81.8	40.1	29.6	59.5	135.6	23.6
Wrong	81.9	40.1	29.7	59.6	135.7	23.6
OnlyOne	81.7	40.1	29.6	59.5	135.3	23.6

set. Practically, we change the value of ϵ from 0.0 to 0.3 with an interval of 0.05. As shown in Figure 6, the AIC model performs the best when the ϵ comes to 0.2. Therefore, we set $\epsilon = 0.2$ as the confidence threshold for the causal dynamics calibration stage.

Effect of Mask Selection Strategy. In our future context modeling framework, for each generated caption pair, we adopt causal dynamics calibration to transfer the knowledge of the NAIC model into the AIC model only on the masked word set S_m. The set is determined by masking words whose AIC-predicted probabilities of the corresponding ground truths are lower than a pre-set threshold ϵ. It is natural to question if exists other masked word selection patterns and how they perform. Therefore, we further investigate the following four variant masking methods:

- ***Random***: For the given sentence S, randomly select k words to be masked and input to the teacher NAIC model to conduct causal dynamics calibration accordingly.
- ***Highest***: As a contrast, we mask the generated words from the AIC model whose predicted probabilities of the ground truth words are higher than the preset threshold ϵ.
- ***Wrong***: Since the ground-truth labels are given, we try to mask the words where the highest probability predictions of the AIC model are different from the corresponding labels.
- ***OnlyOne***: In this variant, to illustrate the necessity of selectively distilling knowledge on a portion rather than all of the target words, we generate NAIC-predicted probability distributions for all target words. As an extreme case, we

Table 5: The percentage of words within each probability interval on the training set.

Model	[0,0.1)	[0.1,0.2)	[0.2,0.3)	[0.3,0.4)	[0.4,0.5)	[0.5,1)
Transformer	26.21	5.13	3.12	4.56	6.14	54.84
FutureCap	25.92	4.78	3.01	4.24	5.92	56.13
Δ	-0.29	-0.35	-0.11	-0.32	-0.22	+1.29

iteratively and only mask one word at once with the given image and residual sentence as input to the NAIC model.

The evaluation results for different masked word selection strategies are presented in Table 4. We can observe that: 1) For "Random" and "Highest" masking strategies, both variants are inferior to our threshold-based causal dynamics calibration method. In particular, the results of "Highest" indicate that conducting dynamics calibration on the confident words is less effective, resulting in a decrease of 0.6 CIDEr score. Meantime, the heuristic selection of masked words is necessary rather than randomness; 2) The result of "Wrong" is lower than our approach. It may be due to the distribution difference between low confidence and incorrect predicted words. 3) "OnlyOne" represents one approach to generating NAIC-predicted probability distributions for all target words iteratively. It is also reasonable for "OnlyOne" to obtain a worse performance since some words can be easy to generate conditioned on local context and over-calibration hold some side effects. All these results demonstrate that it is crucial for the AIC model to exploit the global context on its unconfident words. At the same time, a more advanced and learnable selection strategy may contribute to better captioning performance.

Influence on Model Confidence Distribution. As the prior experiments show one drawback of conventional image captioning lies in the low confidence to correct words, here we also investigate the change of image captioning model confidence with respect to ground-truth words on the MS COCO training set with future context modeling. We list the percentage of words within each interval, in terms of AIC-predicted probability in Table 5. As the probability higher than 0.5 must be the maximum across the vocabulary, we chunk 0.5-1.0 as a high-confidence interval while the others are subdivided into low-confidence intervals. According to the evaluation results, it is obvious that the number of words in low-confidence intervals drops with FutureCap. For instance, the number of words located in [0.1, 0.2) becomes 0.35% fewer. It indicates that our FutureCap model becomes more confident about the ground-truth words given the accurate context.

4.4 Case Study

In order to qualitatively show the effectiveness of future context modeling, we showcase several generated image description results from conventional Transformer with mesh-memory and our FutureCap model, as well as the human-annotated ground-truth sentences (GT) in Figure 7. Generally, it is easy to see that both approaches are able to produce linguistically coherent descriptions. Nevertheless, when examining the fine-grained image content, our future information incorporated method produces more accurate and fluent descriptive sentences by exploiting global information

GT: *a group of people preparing food in a line*

Transformer: *a group of people working together*

FutureCap: *a group of people work green vegetable together*

GT: *a black man sitting on a bench behind a personal bicycle*

Transformer: *a man sitting on a bench on a bicycle*

FutureCap: *a man sitting on a bench next to a bicycle*

GT: *a little girl in a public bathroom for kids*

Transformer: *a girl in blue dress standing next to a toilet*

FutureCap: *a girl in blue dress standing next to a toilet in a bathroom*

Figure 7: Case studies of original Transformer and our FutureCap model, coupled with the corresponding ground truth sentences (GT).

for different word predictions. For example, plain Transformer generates the phrase *on a bicycle* which is inconsistent with the visual relationship for the second image, while the words *next to a bicycle* in our model depicts more precisely. This again confirms the advantage of capturing global context when applying the proposed FutureCap method.

5 RELATED WORKS

Image Captioning. In recent years, a large number of neural systems have been proposed for the image captioning task [3, 9, 16, 22, 24, 40, 53, 58]. The state-of-the-art approaches depend on the encoder-decoder framework to translate the image into a descriptive sentence. Specifically, the encoder network computes visual representations for the image and the decoder network generates a target sentence based on the visual representations. To allow more effective use of the visual representations, a series of attention models have been proposed and achieved great success in multiple sequence-to-sequence learning tasks [4, 38]. In recent years, Transformer-based architectures [9, 14, 17, 26, 32, 40, 59] are introduced to replace conventional RNN, achieving new state-of-the-art performances. On the other hand, lots of mask-based non-autoregressive decoding methods are studied for inference acceleration with a global perspective [13, 15, 17, 18, 20]. However, as far as we are concerned, improving the original language decoding with supervised future information from the NAIC decoder has never been studied in image captioning, which pushes forward our exploration in this paper.

Training Procedure. Training strategy for image captioning models usually follows the word-level cross-entropy paradigm from left to right. This was later combined with a fine-tuning phase based on the application of the REINFORCE method, to allow use as optimization objectives captioning metrics directly [36, 46], boosting the final performance. As a strategy to improve both training phases, in [25] it is proposed to exploit a teacher model trained on image attributes to generate additional supervision signals for the captioning model. These are in the form of soft labels, which the captioning model has to align within the cross-entropy phase, and re-weighting of the caption words to guide the fine-tuning phase. [5] improve the quality with the interaction of two interconnected language models that learn from each other. Additional improvement to the performance of recent self-attention-based image captioning approaches is due to the use of large-scale vision-and-language pre-training [8, 33, 43, 62, 65], which can be done on noisy and weakly annotated image-text pairs, also exploiting pre-training losses different from cross-entropy, such as the masked word loss [62]. Different from all previous methods, our approach is based on the assistance of an additional non-autoregressive image captioning model that is trained with the muti-task learning and dynamic distribution calibration, without changing the internal model architecture or relying on a prior large-scale pre-training model.

Future Information Incorporation. There are numerous works [7, 11, 12, 27, 39, 42, 45] dived to exploit the future information to boost the performance for sequence-to-sequence learning. However, their modelings are different from ours. Specifically, to exploit the future information, [7] adopt a fine-tuned BERT [10] to encode the words that will be generated in the future to acquire the global cost and then exploited as extra supervision to guide the current word generation. [1, 64] employ an extra teacher network to help the neural machine translation model capture global information with knowledge distillation. For [11, 42, 45], given the previous history, in addition to the current target, they further predict the future words, *i.e.*, [11, 42] one more step ahead, and [45] the rest of the sequence. [39] only consider the current target to model the future information and [48, 50, 55] regularize the right-to-left generation, while we directly leverage the effective knowledge to enhance the modeling of the future information. The most similar work is [67], both of us point the importance of bi-direction context and employ it for improved image captioning, [67] introduce a compact directional transformer to parallel decoding while we devise causal dynamics calibration without extra parameters.

6 CONCLUSION

In this paper, we focus on making a conventional image captioning model to effectively exploit the global context without any extra inference cost. Specifically, we resort to the mask-based non-autoregressive decoder for future information modeling during training. Specifically, we introduce multi-task learning to benefit the AIC model by sharing its visual encoder with an auxiliary NAIC. Next, we explore distilling a teacher NAIC model by training the AIC student model to capture the causal dynamics for unconfident words. Experimental results on the MS COCO dataset show that our future information incorporation framework can significantly improve the captioning performance. More importantly, no additional carefully designed network is needed and only the original image captioning model is involved during inference.

REFERENCES

[1] Xi Ai and Bin Fang. 2021. Almost Free Semantic Draft for Neural Machine Translation. In *Proc. NAACL*. 3931–3941.

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In *Proc. ECCV*. 382–398.

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *Proc. IEEE CVPR*. 6077–6080.

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[5] Manuele Barraco, Matteo Stefanini, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. 2022. CaMEL: Mean Teacher Learning for Image Captioning. *arXiv preprint arXiv:2202.10492* (2022).

[6] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* (2015).

[7] Yen-Chun Chen, Zhe Gan, Yu Cheng, Jingzhou Liu, and Jingjing Liu. 2020. Distilling Knowledge Learned in BERT for Text Generation. In *Proc. ACL*. 7893–7905.

[8] Marcella Cornia, Lorenzo Baraldi, Giuseppe Fiameni, and Rita Cucchiara. 2021. Universal Captioner: Long-Tail Vision-and-Language Model Training through Content-Style Separation. *arXiv preprint arXiv:2111.12727* (2021).

[9] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-Memory Transformer for Image Captioning. In *Proc. IEEE CVPR*. 10578–10587.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[11] Chaoqun Duan, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Conghui Zhu, and Tiejun Zhao. 2021. Modeling future cost for neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 770–781.

[12] Zhengcong Fei. 2020. Actor-critic sequence generation for relative difference captioning. In *Proc. ICMR*. 100–107.

[13] Zhengcong Fei. 2020. Iterative Back Modification for Faster Image Captioning. In *Proc. ACM MM*. 3182–3190.

[14] Zhengcong Fei. 2021. Memory-Augmented Image Captioning. In *Proc. AAAI*, Vol. 35. 1317–1324.

[15] Zhengcong Fei. 2021. Partially non-autoregressive image captioning. In *Proc. AAAI*, Vol. 35. 1309–1316.

[16] Zhengcong Fei. 2022. Attention-Aligned Transformer for Image Captioning. In *Proc. AAAI*. 3931–3941.

[17] Zheng-cong Fei. 2019. Fast image caption generation with position alignment. *arXiv preprint arXiv:1912.06365* (2019).

[18] Junlong Gao, Xi Meng, Shiqi Wang, Xia Li, Shanshe Wang, Siwei Ma, and Wen Gao. 2019. Masked non-autoregressive image captioning. *arXiv preprint arXiv:1906.00717* (2019).

[19] Jiuxiang Gu, Jianfei Cai, Gang Wang, and Tsuhan Chen. 2018. Stack-captioning: Coarse-to-fine learning for image captioning. In *Proc. AAAI*, Vol. 32.

[20] Longteng Guo, Jing Liu, Xinxin Zhu, Xingjian He, Jie Jiang, and Hanqing Lu. 2020. Non-autoregressive image captioning with counterfactuals-critical multi-agent learning. *arXiv preprint arXiv:2005.04690* (2020).

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proc. IEEE CVPR*. 770–778.

[22] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. Image Captioning: Transforming Objects into Words. In *Proc. NIPS*. 11135–11145.

[23] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* 2, 7 (2015).

[24] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. In *Proc. IEEE ICCV*. 4634–4643.

[25] Yiqing Huang and Jiansheng Chen. 2020. Teacher-Critical Training Strategies for Image Captioning. *arXiv preprint arXiv:2009.14405* (2020).

[26] Jiayi Ji, Yunpeng Luo, Xiaoshuai Sun, Fuhai Chen, Gen Luo, Yongjian Wu, Yue Gao, and Rongrong Ji. 2021. Improving image captioning by leveraging intra-and inter-layer global representation in transformer network. In *Proc. AAAI*, Vol. 35. 1655–1663.

[27] Kushal Kafle and Christopher Kanan. 2016. Answer-type prediction for visual question answering. In *Proc. IEEE CVPR*. 4976–4984.

[28] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proc. IEEE CVPR*. 3128–3137.

[29] Mahmoud Khademi and Oliver Schulte. 2018. Image caption generation with hierarchical contextual visual spatial attention. In *Proc. IEEE CVPR*. 1943–1951.

[30] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[31] Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proc. ACL Workshop*. 228–231.

[32] Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. 2019. Entangled Transformer for Image Captioning. In *Proc. IEEE ICCV*. 8928–8937.

[33] 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 *Proc. ECCV*. Springer, 121–137.

[34] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of summaries. *Proc. ACL Workshops*, 74–81.

[35] Fenglin Liu, Xuancheng Ren, Xian Wu, Shen Ge, Wei Fan, Yuexian Zou, and Xu Sun. 2020. Prophet Attention: Predicting Attention with Future Attention for Improved Image Captioning. In *Proc. NIPS*. 1–12.

[36] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. 2017. Improved image captioning via policy gradient optimization of spider. In *Proc. IEEE CVPR*. 873–881.

[37] Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, Yongjian Wu, Feiyue Huang, Chia-Wen Lin, and Rongrong Ji. 2021. Dual-level Collaborative Transformer for Image Captioning. In *Proc. AAAI*, Vol. 35. 2286–2293.

[38] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* (2015).

[39] Chih-Yao Ma, Yannis Kalantidis, Ghassan AlRegib, Peter Vajda, Marcus Rohrbach, and Zsolt Kira. 2020. Learning to generate grounded visual captions without localization supervision. In *Proc. ECCV*. Springer, 353–370.

[40] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. 2020. X-linear attention networks for image captioning. In *Proc. IEEE CVPR*. 10971–10980.

[41] Kishore Papineni, Salim Roukos, Todd Ward, and Wei Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proc. ACL*. 311–318.

[42] Yu Qin, Jiajun Du, Yonghua Zhang, and Hongtao Lu. 2019. Look back and predict forward in image captioning. In *Proc. IEEE CVPR*. 8367–8375.

[43] 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 *Proc. ICML*. PMLR, 8748–8763.

[44] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proc. NIPS*. 91–99.

[45] Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. 2017. Deep reinforcement learning-based image captioning with embedding reward. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 290–298.

[46] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-Critical Sequence Training for Image Captioning. In *Proc. IEEE CVPR*. 1179–1195.

[47] Fawaz Sammani and Mahmoud Elsayed. 2019. Look and modify: Modification networks for image captioning. *arXiv preprint arXiv:1909.03169* (2019).

[48] Fawaz Sammani and Luke Melas-Kyriazi. 2020. Show, edit and tell: a framework for editing image captions. In *Proc. IEEE CVPR*. 4808–4816.

[49] Zeliang Song, Xiaofei Zhou, Zhendong Mao, and Jianlong Tan. 2021. Image captioning with context-aware auxiliary guidance. In *Proc. AAAI*, Vol. 35. 2584–2592.

[50] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. 2021. From show to tell: A survey on image captioning. *arXiv preprint arXiv:2107.06912* (2021).

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Proc. NIPS*. 5998–6008.

[52] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proc. IEEE CVPR*. 4566–4575.

[53] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proc. IEEE CVPR*. 3156–3164.

[54] Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel. 2016. Image captioning with deep bidirectional LSTMs. In *Proc. ACM MM*. 988–997.

[55] Cheng Wang, Haojin Yang, and Christoph Meinel. 2018. Image captioning with deep bidirectional LSTMs and multi-task learning. *ACM Transactions on Multimedia Computing, Communications, and Applications* 14, 2s (2018), 1–20.

[56] Li Wang, Zechen Bai, Yonghua Zhang, and Hongtao Lu. 2020. Show, Recall, and Tell: Image Captioning with Recall Mechanism.. In *Proc. AAAI*. 12176–12183.

[57] Yiren Wang, Yingce Xia, Fei Tian, Fei Gao, Tao Qin, Cheng Xiang Zhai, and Tie-Yan Liu. 2019. Neural machine translation with soft prototype. *Proc. NIPS* 32 (2019).

[58] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proc. ICML*. 2048–2057.

[59] Xu Yan, Zhengcong Fei, Zekang Li, Shuhui Wang, Qingming Huang, and Qi Tian. 2021. Semi-Autoregressive Image Captioning. In *Proc. ACM MM*. 2708–2716.

[60] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring Visual Relationship for Image Captioning. In *Proc. ECCV*. 684–699.

[61] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2017. Boosting image captioning with attributes. In *Proc. IEEE CVPR*. 4894–4902.

[62] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proc. IEEE CVPR*. 5579–5588.

[63] Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, Yiyi Zhou, Yongjian Wu, Feiyue Huang, and Rongrong Ji. 2021. RSTNet: Captioning with Adaptive Attention on Visual and Non-Visual Words. In *Proc. IEEE CVPR*. 15465–15474.

[64] Chulun Zhou, Fandong Meng, Jie Zhou, Min Zhang, Hongji Wang, and Jinsong Su. 2022. Confidence Based Bidirectional Global Context Aware Training Framework

[65] for Neural Machine Translation. *arXiv preprint arXiv:2202.13663* (2022).

[65] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. In *Proc. AAAI*, Vol. 34. 13041–13049.

[66] Long Zhou, Jiajun Zhang, and Chengqing Zong. 2019. Synchronous bidirectional neural machine translation. *Transactions of the Association for Computational Linguistics* 7 (2019), 91–105.

[67] Yuanen Zhou, Zhenzhen Hu, Daqing Liu, Huixia Ben, and Meng Wang. 2022. Compact Bidirectional Transformer for Image Captioning. *arXiv preprint arXiv:2201.01984* (2022).