Exploring Non-Autoregressive Image Captioning: Patterns and Semantics

Anonymous ACL submission

Abstract

In the realm of image captioning (IC), learning sentence pattern and semantics plays a crucial role. The reason why this aspect has not received enough attention before is that the prevailing IC models utilize the autoregressive IC (AR-IC) paradigm which operates in a word-by-word manner. In this paradigm, coherence and fluency with the previous text are prioritized during word generation, without special considerations for the sentence pattern. While effective, the AR-IC approaches pose inherent challenges for real-time applications due to their time-consuming nature during inference. Unlike the AR-IC counterparts, non-autoregressive IC (NAR-IC) models necessitate simultaneous inference of all words in a caption. However, the existing NAR-IC models have been met with the hurdle of reduced effectiveness in comparison to their autoregressive counterparts. It is largely because they follow the AR-IC approach, neglecting the influence of patterns and semantics on NAR-IC. Considering that the dependency on preceding and following words is eliminated during NAR-IC generation, it becomes crucial to consider the sentence pattern to guide word generation. In this paper, we reconsider the impact of sentence patterns and semantics in NAR-IC training. We delve into NAR-IC and provide tips and tricks for training NAR-IC models, which include knowledge distillation, label selection, image pre-fusion, and NAR+AR enhancement. By meticulously examining the impact of these components on model performance, we achieve the state-of-the-art performance with a single-step generation. This paper aims to provide valuable strategies for those aiming to advance NAR-IC models. Our code is provided in Supplementary materials.

1 Introduction

Image captioning (IC) has received substantial attention in recent years, which aims to provide descriptive narratives for input images. Autoregressive IC (AR-IC) models, which generate captions word-by-word, have been a prominent approach in this domain (Vinyals et al., 2015; Xu et al., 2015; Jiang et al., 2018). Nevertheless, AR-IC models face a significant constraint related to both training and inference speed. As depicted in Figure 1 (a), AR-IC employ masked sentences to replicate the current state of the sentence and predict the next word. This limitation becomes more pronounced when considering resource-constrained devices or real-time applications (Gu and Tan, 2022). After the introduction of Transformers (Vaswani et al., 2017), the parallelism they offer provides an opportunity to represent a departure from the conventional word-by-word generation pattern, which is non-autoregressive IC (NAR-IC). However, the simultaneous generation of all words in the sentence leads to issues such as word repetition and sentence disorder (Ran et al., 2021; Gu and Kong, 2020; Xiao et al., 2023). Two solutions have emerged to address this issue. One involves increasing the number of iterations, such as through the refinement of generated captions (Lee et al., 2018; Ghazvininejad et al., 2019; Fei et al., 2023), and diffusion models (Zhu et al., 2022; Luo et al., 2023), as Figure 1 (b) shows. However, the refinement or diffusion process is inherently iterative and non-parallelizable, which poses an efficiency challenge when improving the caption quality. To alleviate this limitation, recent researchers focus on elevating caption quality while simultaneously preserving efficiency, especially within a single-step generation (Guo et al., 2020; Yu et al., 2023). As is shown in Figure 1 (d), these NAR-IC methods with a single inference rely solely on the input image. However, due to the nature of IC, the output sequence is sequential while the input image is not. Aligning the non-sequential image patches with the sequential words is challenging.

Researchers in cognitive science (Hale et al., 2018; Ryskin and Nieuwland, 2023) provide inter-

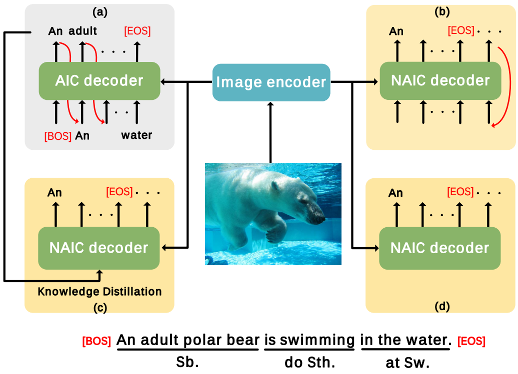

Figure 1: Comparison of AR-IC and NAR-IC models, where panel (a) represents an AR-IC model, and three different types of NAR-IC models in panels (b), (c), and (d). The red line represents the iteration in inference.

esting findings that humans prioritize considering the sentence structure when producing sentences. It differs from the traditional word-by-word approach of AR-IC methods, but shares similarities with the NAR-IC generation (Yang et al., 2019; Fisch et al., 2020). Inspired by this insight, the process of NAR-IC can primarily focus on exploring the sentence pattern and subsequently filling in the semantics into this sentence pattern. From the perspective of sentence composition, the sentence pattern comprises sequential features, and the semantics emerge once these sequential features are decoupled, *e.g.*, after determining the sentence pattern " Sb. do Sth. at Sw.", the model only needs to find useful information from the image to fill in the placeholders these "some".

To summarize the temporal information into a sentence pattern, the majority of current NAR-IC models utilize knowledge distillation (Guo et al., 2020; Yu et al., 2023) from AR-IC models. The sentences generated by AR-IC models instruct the NAR-IC model to follow their patterns, as depicted in Figure 1 (c). However, knowledge distillation presents two primary drawbacks. Firstly, it requires an additional training phase for a strong AR-IC model to generate advanced labels. Secondly, due to the performance constraints of the AR-IC model, the labels derived from knowledge distillation may not consistently be accurate.

In this paper, we explore the solutions for learning patterns and semantics. Specifically, label selection serves as an alternative to knowledge distillation for learning the sentence pattern. It involves selecting optimal image-caption pairs from the ground-truth annotations. The most advantage of label selection over knowledge distillation is that the labels distilled by AR-IC may not always be consistent with the image. Additionally, it does not necessitate the additional training of a strong AR-IC model. Instead, it only requires a weak AR-IC model or independent evaluation metrics (*e.g.*, CLIP score (Hessel et al., 2021)) to select labels with proper patterns from the ground-truth annotations, thereby simplifying the training process. It significantly reduces the difficulty of NAR-IC training, as it reduces the dependency between contextual words and focuses on extracting effective information from the image. Regarding the semantic part, given that it has decoupled the temporal nature of the sentence, it becomes crucial to extract effective information from the image. Therefore, an image pre-fusion module is proposed to fuse the image feature into the decoder. It allows more image information to be mapped to the corresponding part of the generated sentence. Besides, we introduce a unified architecture to train in a NAR+AR paradigm, which allows NAR-IC to learn more structure modalities and semantics from the AR-IC training. In detail, NAR-IC is first trained to enable the model to learn specific patterns and overcome the temporal dependencies. Following this, AR-IC is integrated into the unified architecture to further improve the word semantics in NAR-IC and boost the performance.

In summary, our research has yielded useful strategies for enhancing the effectiveness of the NAR-IC model. The key findings from our study, along with detailed results available in Table 1, are as follows:

- **Knowledge Distillation**: Employing knowledge distillation techniques to transfer knowledge from well-performing AR-IC models to boost the performance of the NAR-IC model;
- **Effective Label Selection**: Leveraging the outcomes of existing caption models to select proper patterns from the ground-truth annotations, thus the preferred image-caption pairs are obtained;
- **Image pre-fusion in Decoder**: Enhancing the connection between images and captions by incorporating image features into the decoder;
- **NAR+AR Training Enhancement**: Implementing a NAR+AR training approach within a shared architecture to further improve the performance.

Through the application of these techniques, our approach attains state-of-the-art performance among

NAR-IC models in a single-step inference, all while maintaining the efficiency characteristic of NAR-IC. The addition of these strategies does not bring additional computation in inference process. While some of these methods have demonstrated effectiveness in prior work, we have conducted comprehensive analyses and experiments to thoroughly explore their impact.

2 Related Works

2.1 Image captioning (IC)

The combination of CNNs for image feature extraction and RNNs for language modeling, introduced by Vinyals *et al.* (Vinyals et al., 2015), has paved the way for end-to-end trainable models capable of generating coherent and contextually relevant captions. Within this architecture, Xu *et al.* (Xu et al., 2015) proposed the attention mechanisms, enabling models to focus on different regions of an image while generating captions. Anderson *et al.* (Anderson et al., 2018) proposed Up-Down model which employed a bottom-up mechanism to align the object regions to the generated words. In recent years, the advent of Transformer-based models (Vaswani et al., 2017) has reshaped the landscape of IC (Huang et al., 2019; Li et al., 2019; Wang et al., 2022; Zhou et al., 2020). For example, Wang *et al.* (Wang et al., 2022) applied Swin Transformer (Liu et al., 2021) for both image encoder and language decoder, benefiting from its unified architecture. The availability of large-scale pre-trained models (Li et al., 2020; Zhang et al., 2021; Li et al., 2022, 2023) has also benefited IC as a downstream task, leading to improvements in caption quality. However, it is noteworthy that the aforementioned models typically follow the AR approach for caption generation, which necessitates substantial computational resources and introduces latency for both training and inference.

2.2 Non-autoregressive (NAR) decoding

Unlike AR decoding, which generates text word by word, NAR models produce the entire sequence in a single inference, making it more efficient during inference. As far as we know, Gu *et al.* (Gu et al., 2018) were among the first to introduce NAR text generation using Transformer-based architectures, allowing for parallel decoding of text sequences. While NAR text generation holds promise for efficient text production, challenges remain, for example, under or over generation, incoherent sentences (Gu and Tan, 2022). Attempts have been made to overcome these issues. For example, Fertility predictor (Ran et al., 2021; Gu et al., 2018) was proposed to predict the length of the generated sentences. Continuous VAEs (Shu et al., 2020) trained a Gaussian prior on each words. Other approaches involve SemiAR, generating text phrase by phrase (Lample et al., 2018; Qi et al., 2020). However, this approach still represents a trade-off between time efficiency and performance.

Recent research has explored NAR image captioning models (Gao et al., 2019; Fei, 2019; Zhu et al., 2022; Yu et al., 2023; Guo et al., 2020; Deng et al., 2020). For example, Gao *et al.* (Gao et al., 2019) introduced NAR to image captioning predicting masked words in parallel. Zhu *et al.* (Zhu et al., 2022) introduced discrete diffusion into NAR-IC, which achieved excellent results. However, because the diffusion model requires multiple refinements, it does not offer an advantage in terms of efficiency. To further accelerate the inference speed, efforts were made to improve the performance of a single-inference NAR-IC model. Guo *et al.* (Guo et al., 2020) introduced reinforcement learning and sequence-level knowledge distillation. On the other hand, Liu *et al.* (Yu et al., 2023) used the image feature as a decoder input, which significantly improved the quality of NAR output in a single inference.

Despite the promise of NAR-IC, generating a high-quality caption with higher time efficiency, especially in a single-step inference, remains a challenge. Furthermore, effectively utilizing high-performance AR pre-trained models within the NAR framework is a crucial aspect of this endeavor.

3 Methods

Given an image I, the caption $\boldsymbol{Y}$ is generated by a captioning model with its parameters θ. This caption can be decomposed into the sentence pattern part $\boldsymbol{Y_p}$ and the semantics part $\boldsymbol{Y_s}$, respectively:

$$p(\boldsymbol{Y}|I;\theta) = p(\boldsymbol{Y_p}|I;\theta)p(\boldsymbol{Y_s}|I;\theta), \quad (1)$$

where $\boldsymbol{Y_p}$ and $\boldsymbol{Y_s}$ are assumed to be conditionally independent.

3.1 Knowledge distillation & Label selection

Take MSCOCO (Lin et al., 2014) for example, each image is associated with five human-annotated captions. By observing these captions, we find that they exhibit various sentence patterns in describing

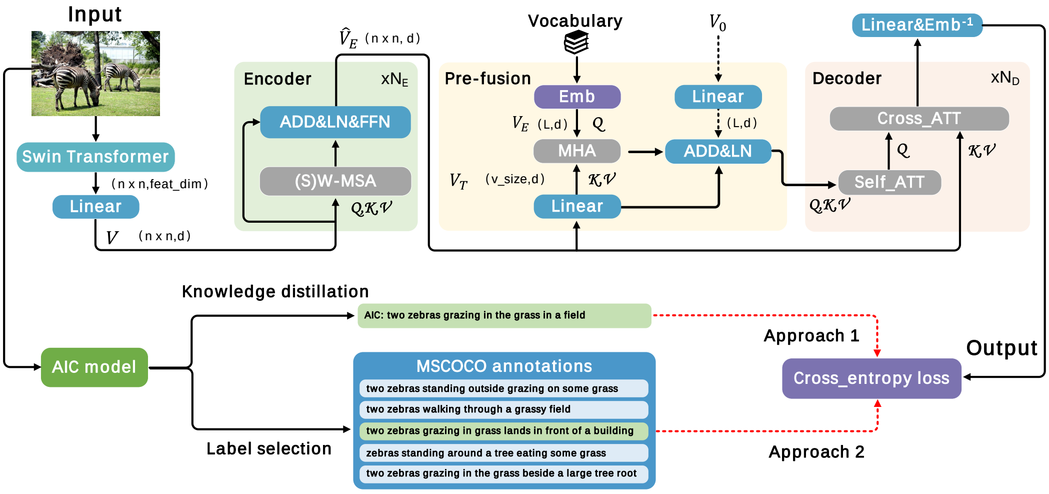

Figure 2: The architecture of our NAR-IC model. In this illustration, "Emb" and "Emb^{-1}" denote the word embedding function and its inverse function. "N_E" and "N_D" refer to the number of encoder and decoder layers, respectively. "feat_dim" represents the feature dimension, and "d" is the embedding size of our model. The black dotted line indicates that it only takes effect in AR training mode. The red dotted line denotes two alternative approaches for choosing labels: either through knowledge distillation (approach 1) or label selection (approach 2).

the same image. Besides, previous works and our experiments have demonstrated that randomly selected sentences exhibit diverse patterns, which are not beneficial for NAR (Guo et al., 2020; Yu et al., 2023; Deng et al., 2020). Since the sentence structure generated by the AR-IC models is relatively uniform, previous NAR-IC models have adopted knowledge distillation.

However, the NAR-IC model implemented through knowledge distillation heavily relies on the quality of the AR-IC model. Thus, we propose an alternative approach for knowledge distillation: label selection, which obtains high-quality annotations from the ground-truth labels. This method involves selecting labels from the ground-truth annotations based on their similarity to the AR-IC-generated results. To be specific, it employs a pre-trained AR-IC model to generate sentences corresponding to the images. Subsequently, a comparison is made between the sentences generated by the AR-IC model and the existing MSCOCO annotations. The labels with the highest similarity metrics are selected to create the training set. Additionally, we investigate the use of other individual evaluation metrics, such as the CLIP score (Hessel et al., 2021), to assess the quality of these labels. As Figure 2 shows, knowledge distillation and label selection act as mutual substitutes.

3.2 Image Pre-fusion

Since sequential pattern $\boldsymbol{Y_p}$ is decoupled, enhancing the connection between images and generated sentences becomes essential to learning the semantic part $\boldsymbol{Y_s}$ in Eq. (1). To achieve this, we employ a linear layer $\mathcal{L}$ to map image features onto the sentence, enabling the model to more effectively extract relevant information from the image:

$$\hat{V_D} = \mathbf{LN}(\mathcal{L}(V_E) + \mathbf{MHA}(\mathcal{L}(V_E), V_T, V_T)), \quad (2)$$

where V_E and V_T represents the embedding vectors of the image and the vocabulary, "MHA" denotes the multi-head attention (Vaswani et al., 2017), "LN" represents the layernorm layer. It pre-fuses the image features as an integral part of the input of the decoder.

Unlike the conventional approach of initializing the decoder with a sentence replete with [MASK] tokens, this modified decoder initialization method leverages the image features through an MHA mechanism. The subsequent stages closely resemble typical encoder-decoder models, commencing with self-attention on the input of the decoder, followed by cross-attention with the image feature:

$$\begin{aligned} \textbf{Self\_ATT: } & V_D = \mathbf{LN}(\hat{V_D} + \mathbf{MHA}(\hat{V_D}, \hat{V_D}, \hat{V_D})); \\ \textbf{Cross\_ATT: } & \hat{V_S} = \mathbf{LN}(V_D + \mathbf{MHA}(V_D, V_E, V_E)), \quad (3) \\ & V_S = \mathbf{LN}(\hat{V_S} + FFN(\hat{V_S})), \end{aligned}$$

where "FFN" represents the feed-forward layer. Consequently, the conditional probabilities of the sentence are calculated as:

$$p(\boldsymbol{Y}) = LP(V_S), \quad (4)$$

where "LP" denotes the linear projection function, responsible for mapping the feature to the distribution of word sequences. Therefore, this modified NAR-IC decoder architecture seamlessly integrates image features into the sentence generation process, enhancing contextual dependencies and improving language fluency. Thus, the image features are fused into $\boldsymbol{Y_s}$ before cross-attention calculation.

3.3 NAR+AR enhancement

In the context of the model architecture outlined in Section 3.2, the NAR-IC model exhibits limitations in terms of contextual dependencies, resulting in issues related to language fluency. In an effort to bridge the gap between AR-IC and NAR-IC models while maintaining a unified architecture, a modification is proposed to Eq. (2):

$$\hat{V_D'} = \mathbf{LN}(V_E + \mathbf{MHA}(V_E, V_T, V_T) + V_0), \qquad (5)$$

where V_0 represents the feature embedding corresponding to the last state of the sentence. During the training process in NAR mode, V_0 remains consistently set to zero, ensuring that the training regimen remains unaffected by this modification. Conversely, during training in AR mode, the inclusion of the previous state of the sentence is taken into account through V_0. You can refer to Figure 2 for the architecture of our model.

By alternating training the NAR and AR modes, structured semantics are implicitly transferred to the NAR models. This approach allows for a seamless transition between AR and NAR paradigms within a unified architecture, fostering an improved capacity to capture context and enhance the fluency of generated language. This bridging mechanism thus paves the way for a more versatile and context-aware image captioning model.

4 Experiments

4.1 Implementation

Following the previous IC models (Anderson et al., 2018; Huang et al., 2019; Wang et al., 2022; Yu et al., 2023), our model is trained and evaluated on the MSCOCO dataset (Lin et al., 2014), which contains 123,287 images (113,278/5000/5000 for training/validation/testing in Karpathy split (Karpathy and Fei-Fei, 2015)). Each image has 5 corresponding annotations. Consistent with the most IC models, our vocabulary contains 9487 common words. We set the maximum sentence length L to 16, the embedding size of the model d to 512, the number of the encoder and decoder layers N_E and N_D to 3, and the number of Transformer heads h to 8. We apply four widely used metrics to evaluate the quality of the generated captions: BLEU (Papineni et al., 2002), METEOR (Agarwal and Lavie, 2007), ROUGE-L (ROUGE, 2004), and CIDEr (Vedantam et al., 2015), abbreviated as B, M, R, and C, respectively. More training details are listed in Supplementary materials.

4.2 Ablation Studies

In Table 1, we present the results of our extensive ablation experiments conducted to validate the effectiveness of the strategies discussed. Additionally, the results for AR models, denoted as A1 and A2, are included to provide a comprehensive basis for comparison.

The effect of image pre-fusion. The results of D1-D4 underscore the importance of incorporating image features as an integral component of the decoder within a single inference. This modification significantly influences the quality of the generated captions.

The effect of label selection. Notably, under the Transformer (L1) and Swin (L2) architectures, the random selection of MSCOCO labels yields results slightly better than using the entire set of labels on AR-IC models. This observation suggests that learning certain sentence pattern within the MSCOCO dataset might be conducive for effectively training Transformer-based image captioning models. Consequently, the exploration of methods for selecting relevant and informative labels in a NAR-IC model is warranted. Furthermore, by comparing the results of L5-L8, we observe that using CIDEr (L8) and ROUGE (L7) metrics leads to better performance. Additionally, introducing individual metrics such as the CLIP (Hessel et al., 2021) score (L9), is also proved to be effective. Our experiments highlight the significant influence of different evaluation metrics on the overall model performance, underscoring the importance of selecting and utilizing appropriate metrics for label selection in NAR-IC training. We further explore which AR-IC pre-trained model achieves the highest performance. Regardless of whether the classic Transformer architecture (L7-L9) or Swin (L10-L12) is employed, the results are remarkably similar because the labels obtained are almost the same after label selection. This finding also indicates that the Transformer-based models, irrespective of their

Table 1: The performances of the ablation models on Karpathy test split.

No.	Models					Metrics				
	AR	Arch.	Distil.	Lbl Sel.	Img pre-fusion	B@1	B@4	M	R	C
AR Baseline										
A1	✓	Transformer	✗	✗	✗	76.1	33.5	27.8	56.1	114.7
A2	✓	Swin	✗	✗	✗	77.1	47.1	28.5	57.5	120.6
Image pre-fusion										
D1	✗	Transformer	Transformer	✗	✗	49.9	4.8	15.3	29.9	40.0
D2	✗	Swin	Swin	✗	✗	50.2	4.8	15.5	32.0	40.9
D3	✗	Transformer	✗	CLIP score	✗	50.0	4.3	15.5	29.9	40.3
D4	✗	Swin	✗	CLIP score	✗	50.3	4.9	15.5	30.4	40.1
Label Selection										
L1	✓	Transformer	✗	Random	✗	76.9	34.5	28.0	56.7	116.4
L2	✓	Swin	✗	Random	✗	77.4	36.8	28.6	57.4	121.5
L3	✗	Transformer	✗	Random	✓	48.5	12.4	17.8	46.7	60.1
L4	✗	Transformer	✗	Loss	✓	79.1	36.0	28.2	57.0	120.3
L5	✗	Transformer	✗	BLEU	✓	71.4	24.9	23.5	51.9	86.1
L6	✗	Transformer	✗	METEOR	✓	70.2	25.2	23.2	52.3	87.1
L7	✗	Transformer	✗	ROUGE	✓	79.6	37.1	27.9	57.6	122.9
L8	✗	Transformer	✗	CIDEr	✓	79.8	36.9	28.1	57.8	121.5
L9	✗	Transformer	✗	CLIP score	✓	79.9	37.1	28.1	57.9	123.3
L10	✗	Swin	✗	CIDEr	✓	79.8	36.9	28.1	57.8	121.5
L11	✗	Swin	✗	ROUGE	✓	79.9	37.0	28.1	57.8	122.0
L12	✗	Swin	✗	CLIP score	✓	79.9	37.1	28.1	57.9	123.3
Knowledge Distillation										
K1	✓	Transformer	Swin	✗	✗	79.9	37.1	28.0	57.9	123.3
K2	✓	Swin	VinVL	✗	✗	81.1	39.6	29.4	59.0	132.2
K3	✗	Transformer	Transformer	✗	✓	76.4	34.9	27.9	56.2	115.7
K4	✗	Transformer	Swin	✗	✓	79.9	37.1	28.0	57.9	123.3
K5	✗	Transformer	VinVL	✗	✓	79.8	37.0	28.0	57.8	122.5
K6	✗	Swin	Transformer	✗	✓	76.3	34.9	27.9	56.2	115.6
K7	✗	Swin	Swin	✗	✓	79.5	36.5	27.8	57.7	122.4
K8	✗	Swin	VinVL	✗	✓	79.5	37.7	28.2	57.8	123.3
NAR+AR Enhancement										
N1	✗	Swin	✗	CLIP score	✓	79.2	36.0	27.8	57.4	119.2
N2	NAR+AR	Swin	✗	CLIP score	✓	79.6	36.5	28.0	57.7	121.2
N3	AR+NAR	Swin	✗	CLIP score	✓	72.5	28.4	25.4	47.6	100.0
N4	NAR+Mixed(Ours)	Swin	✗	CLIP score	✓	79.9	37.3	28.2	58.1	123.7

specific architecture, exhibit a tendency to generate captions with similar sentence structures. After comparing the results of the models under cross entropy loss, we observe that our NAR-IC model with label selection by CLIP score (L9 and L12) and ROUGE (L7 and L11) has higher performance than the AR-IC models (L1 and L2). Therefore, utilizing a relatively weak AR-IC model to select valuable labels from ground-truth annotations has been proven effective, in the absence of a high-quality AR-IC pre-trained model.

The effect of knowledge distillation. An important insight emerges when we compare the results of label selection and knowledge distillation. This comparison leads to the formulation of an effective training strategy, that is: when a high-quality AR-IC model, such as Swin (Liu et al., 2021; Wang et al., 2022) and VinVL (Zhang et al., 2021) based AR-IC models, is available, applying knowledge distillation proves to be a more effective and efficient strategy. The results of K4 and K5 suggest that they leverage the knowledge and competence of the pre-trained AR-IC model to enhance the performance of the NAR-IC model. Conversely, employing knowledge distillation becomes a less favorable strategy when the pre-trained AR-IC model is relatively weak, such as the classic Transformer structure (K3). It ensures that the labels chosen for training are more representative and beneficial for the non-autoregressive model, compensating for the potential limitations of the AR-IC model. Besides, an intriguing observation emerges from our study regarding knowledge distillation. It appears that knowledge distillation is not overly sensitive

Table 2: Comparison with the SOTA image captioning methods.

Model	B@1	B@4	M	R	C	SpeedUp
AR-IC Models						
AR	76.9	34.5	28.0	56.7	116.4	1.0×
AR(RL)	80.3	38.4	29.0	58.7	128.8	1.0×
PureT (Wang et al., 2022)	77.3	37.0	28.6	57.4	121.4	4.3×
PureT (Wang et al., 2022)(RL)	82.1	40.9	30.2	60.1	138.2	4.3×
SemiAR-IC Models						
PNAIC (Fei, 2021)	79.9	37.5	28.2	58.0	125.2	6.9×
SATIC (Zhou et al., 2021)	80.6	37.6	28.3	58.1	126.2	6.3×
SAIC (Yan et al., 2021)	80.3	38.4	29.0	58.2	127.1	4.1×
NAR-IC Models						
MNAIC (Gao et al., 2019)	75.4	30.9	27.5	55.6	108.1	3.6×
FNAIC (Fei, 2019)	-	36.2	27.1	55.3	115.7	8.2×
LaBert (Deng et al., 2020)	77.4	35.0	27.9	57.0	116.8	9.3×
CMAL-COCO (Guo et al., 2020)	60.7	15.9	18.2	45.9	60.6	13.9×
CMAL-KD (Guo et al., 2020)	78.5	35.3	27.3	56.9	115.5	13.9×
CMAL (Guo et al., 2020) (RL)	80.3	37.3	28.3	58.0	124.0	13.9×
EENAIC-COCO (Yu et al., 2023)	60.2	16.0	17.7	45.5	60.1	37.0×
EENAIC-KD (Yu et al., 2023)	79.7	36.9	27.9	58.0	122.6	37.0×
Ours-KD	79.9	37.3	28.2	58.1	123.7	37.0×
Ours-COCO	80.0	37.2	28.3	58.2	123.6	37.0×
Ours-KD (RL)	80.1	37.3	28.2	58.3	123.9	37.0×
Ours-COCO (RL)	80.3	36.8	28.2	58.3	125.2	37.0×

to the architectural consistency between the teacher model and the student model. Instead, the critical factor influencing the effectiveness of knowledge distillation is the quality of the teacher model. In other words, while having consistent architectures between the teacher and student models can be beneficial, it is not a strict requirement. What truly matters is the capability and performance of the teacher model. For example, despite K4 employing a unified Swin structure in both the teacher and student model, it fails to surpass the performance of K5, which utilizes VinVL as the teacher model and Swin as the student model.

The effect of NAR+AR enhancement. The results indicate that training the NAR model initially and subsequently adding AR training (cf. Eq. (5)) leads to the best overall performance (N4). Moreover, when AR training is conducted first (N3), the model acquires an understanding of the temporal dependencies that are inherent in the autoregressive generation process. However, when the training shifts to NAR mode, it becomes challenging for the model to break free from these learned dependencies. As a consequence, this results in a performance drop in the NAR mode.

4.3 Comparisons with SOTA

In Table 2, we present performance comparisons of our best model with existing methods, including MNAIC (Gao et al., 2019), FNAIC (Fei, 2019), Labert (Deng et al., 2020), CMAL (Guo et al., 2020), and EENAIC (Yu et al., 2023). It is important to note that MNAIC (Gao et al., 2019), FNAIC (Fei, 2019), and Labert (Deng et al., 2020) adopt refinement strategies, which entail a more time-consuming inference process. On the other hand, CMAL (Guo et al., 2020) and EENAIC (Yu et al., 2023), like our model, generate captions within a single inference step, emphasizing efficiency. We present two sets of results: "COCO" where we exclusively utilize selected MSCOCO (Lin et al., 2014) annotations during training (corresponding to L12 in Table 1), and "KD" which signifies the usage of knowledge distillation (corresponding to K8 in Table 1). Additionally, we list some AR-IC and SemiAR-IC models for reference.

We observe that our model achieves the best performance among the NAR models. Besides, we deliver a substantial improvement in inference speed, approximately three times faster. Moreover, the enhancement of "Ours-KD" after reinforcement learning (RL) training is not as pronounced as that seen in "Ours-COCO". The primary reason is that the labels used in knowledge distillation are obtained by the pre-trained AR-IC+RL model. In addition, we compare the results of the models using annotations only from MSCOCO (Lin et al., 2014). Unlike knowledge distillation, which requires a strong AR-IC model to instruct the NAR-IC model, our approach employs a weaker AR-IC model with a CIDEr of 116.4 to select preferred image-caption pairs, ultimately achieving a CIDEr of 123.5. Besides, we observe that methods like CMAL (Guo et al., 2020) and EENAIC (Yu et al., 2023) fail to deliver satisfactory results without knowledge distillation. It indicates the broader applicability and effectiveness of our method. Additionally, it is noteworthy that "Ours-COCO" demonstrates comparable performance to "Ours-KD" even without knowledge distillation. "Ours-COCO" entirely eliminates the influence of knowledge distillation and RL training, resulting in a CIDEr score of 123.6. This score is 0.4 lower than CMAL (Guo et al., 2020) with RL training and 8.1 higher than CMAL without RL.

The results from the MSCOCO online test are also presented in Table 3, where "*" denotes our unofficial submission. Our model attains a comparable performance to early AR-IC methods like SCST (Rennie et al., 2017) and Up-Down (Anderson et al., 2018). This suggests that our NAR-IC model holds the potential to replace the early AR models in terms of performance, all while offering a significant advantage in terms of inference speed. When compared with the models under the cross entropy loss, our method ("Ours-COCO")

Table 3: The scores on the MSCOCO online test server.

Models	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
AR-IC models														
SCST(RL) (Rennie et al., 2017)	78.1	93.7	61.9	86.0	47.0	75.9	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.7
Up-Down(RL) (Anderson et al., 2018)	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(RL) (Huang et al., 2019)	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
PureT(XE)* (Wang et al., 2022)	75.8	93.9	59.1	86.3	45.0	76.0	34.1	41.4	27.7	37.6	55.7	71.7	111.3	114.7
PureT(RL) (Wang et al., 2022)	82.8	96.5	68.1	91.8	53.6	83.9	41.4	74.1	30.1	39.9	60.4	75.9	136.0	138.3
SemiAR-IC models														
PNAIC (Fei, 2021)	80.1	94.4	64.0	88.1	49.2	78.5	36.9	68.2	27.8	36.4	57.6	72.2	121.6	122.0
NAR-IC models														
CMAL(RL) (Guo et al., 2020)	79.8	94.3	63.8	87.2	48.8	77.2	36.8	66.1	27.9	36.4	57.6	72.0	119.3	121.2
EENAIC* (Yu et al., 2023)	79.0	93.8	62.5	85.6	47.5	75.0	35.6	63.9	27.6	36.2	57.1	71.4	115.4	117.5
Ours-KD	79.3	93.9	62.9	86.1	47.9	75.8	35.9	64.8	27.8	36.3	57.3	71.6	116.8	118.8
Ours-COCO	79.2	93.9	62.9	86.2	47.8	76.0	35.9	65.0	27.8	36.5	57.3	71.9	116.7	118.9
Ours-KD(RL)	79.3	94.1	63.3	86.9	48.8	76.4	36.2	65.8	27.9	36.9	57.6	71.9	116.9	118.9
Ours-COCO(RL)	80.0	94.6	63.6	87.5	49.9	79.9	37.8	67.1	28.2	37.3	58.0	72.5	119.5	122.4

achieves 116.8/118.8 on CIDEr c5/c40, which outperforms the AR-IC method PureT(XE). Furthermore, when contrasted with models incorporating RL training, "Ours-COCO" attains the highest performance among the NAR-IC methods and is comparable to the SemiAR (semi-autoregressive) method PNAIC (Fei, 2021). In comparison to the AR-IC methods, we exhibit the closest performance and significantly faster inference speeds.

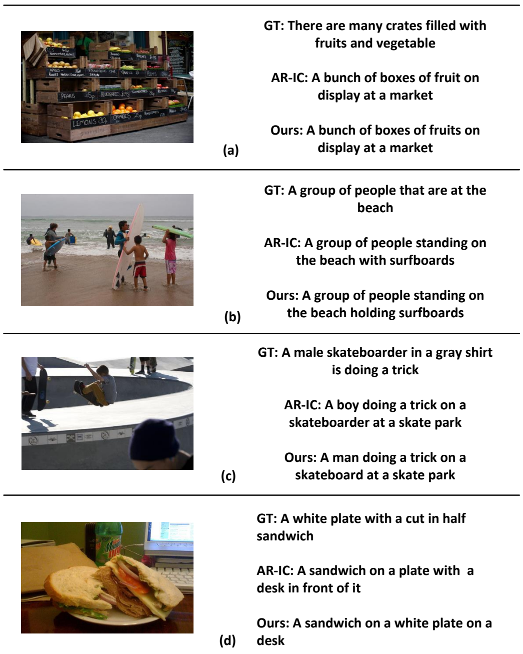

Figure 3: Examples of the Ground-truth captions (GT), generated captions by AR-IC and our model.

4.4 Qualitative Results

The qualitative results are shown in Figure 3. In the first scenario, as depicted in Figures 3 (a), (b), and (c), our NAR-IC model inherits valuable insights from the AR-IC model. Notably, this highlights the advantage of our NAR-IC model in terms of inference speed, as it can achieve comparable results without the sequential word-by-word generation characteristic of AR-IC models. In the second scenario, exemplified in Figure 3 (d), the results of the AR-IC model exhibit inaccuracies in the description, such as the imprecise usage of "in front of". When the AR-IC model predicts the wrong word "in" instead of "on", it tends to subsequently predict "front of", diverging from the ground truth. In this case, our NAR-IC model outperforms the AR-IC model, providing descriptions that better align with the ground truth. This demonstrates the potential of the NAR-IC models in producing more accurate and contextually relevant captions, in addition to their remarkable inference speed.

5 Conclusions

This paper delve into the crucial components of the NAR-IC model, including image pre-fusion, knowledge distillation, label selection, and training policies. We analyze the respective significance and effectiveness of each of these components. These observations highlight the strengths and weaknesses of both NAR-IC and AR-IC models. Leveraging these insights, our NAR-IC method demonstrates the potential to combine the efficiency and quality advantages of both paradigms. Our findings underscore the significance of a thoughtful label selection strategy for NAR-IC models and the utilization of existing AR-IC models. The comprehensive experiments we conduct and the careful exploration of various design choices make a substantial contribution to the field, serving as a strong foundation for future research.

6 Limitations

One limitation lies in the fact that, despite significantly accelerating the speed of inference, our proposed NAR-IC method still lacks significant advantages over traditional AR-IC during the training phase. Besides, our method is proved efficient and effective on MSCOCO dataset. However, the MSCOCO dataset consists of accurately labeled images. Our method requires prior denoising when applying on the dataset with noise. Further studies will aim to train NAR on noisy datasets and expand the training scale.

References

Abhaya Agarwal and Alon Lavie. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. *Proceedings of WMT-08*.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Chaorui Deng, Ning Ding, Mingkui Tan, and Qi Wu. 2020. Length-controllable image captioning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 712–729. Springer.

Zheng-cong Fei. 2019. Fast image caption generation with position alignment. *arXiv preprint arXiv:1912.06365*.

Zhengcong Fei. 2021. Partially non-autoregressive image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1309–1316.

Zhengcong Fei, Mingyuan Fan, Li Zhu, Junshi Huang, Xiaoming Wei, and Xiaolin Wei. 2023. Uncertainty-aware image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 614–622.

Adam Fisch, Kenton Lee, Ming-Wei Chang, Jonathan H Clark, and Regina Barzilay. 2020. Capwap: Captioning with a purpose. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, page 8755–8768.

Junlong Gao, Xi Meng, Shiqi Wang, Xia Li, Shanshe Wang, Siwei Ma, and Wen Gao. 2019. Masked non-autoregressive image captioning. *CoRR*.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Constant-time machine translation with conditional masked language models. *arXiv preprint arXiv:1904.09324*.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. 2018. Non-autoregressive neural machine translation. In *International Conference on Learning Representations*.

Jiatao Gu and Xiang Kong. 2020. Fully non-autoregressive neural machine translation: Tricks of the trade. *arXiv preprint arXiv:2012.15833*.

Jiatao Gu and Xu Tan. 2022. Non-autoregressive sequence generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 21–27.

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*.

John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan R Brennan. 2018. Finding syntax in human encephalography with beam search. *arXiv preprint arXiv:1806.04127*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4634–4643.

Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. 2018. Recurrent fusion network for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 499–515.

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

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. *arXiv preprint arXiv:1804.07755*.

Jason Lee, Elman Mansimov, and Kyunghyun Cho. 2018. Deterministic non-autoregressive neural sequence modeling by iterative refinement. *arXiv preprint arXiv:1802.06901*.

Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. 2022. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*.

Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. 2019. Entangled transformer for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8928–8937.

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

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 *European Conference on Computer Vision*, pages 121–137. Springer.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022.

Jianjie Luo, Yehao Li, Yingwei Pan, Ting Yao, Jianlin Feng, Hongyang Chao, and Tao Mei. 2023. Semantic-conditional diffusion networks for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23359–23368.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. *arXiv preprint arXiv:2001.04063*.

Qiu Ran, Yankai Lin, Peng Li, and Jie Zhou. 2021. Guiding non-autoregressive neural machine translation decoding with reordering information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13727–13735.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024.

Lin CY ROUGE. 2004. A package for automatic evaluation of summaries. In *Proceedings of Workshop on Text Summarization of ACL, Spain*.

Rachel Ryskin and Mante S Nieuwland. 2023. Prediction during language comprehension: what is next? *Trends in Cognitive Sciences*.

Raphael Shu, Jason Lee, Hideki Nakayama, and Kyunghyun Cho. 2020. Latent-variable non-autoregressive neural machine translation with deterministic inference using a delta posterior. In *Proceedings of the aaai conference on artificial intelligence*, pages 8846–8853.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Yiyu Wang, Jungang Xu, and Yingfei Sun. 2022. End-to-end transformer based model for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2585–2594.

Yisheng Xiao, Lijun Wu, Junliang Guo, Juntao Li, Min Zhang, Tao Qin, and Tie-yan Liu. 2023. A survey on non-autoregressive generation for neural machine translation and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

Shitong Xu. 2022. Clip-diffusion-lm: Apply diffusion model on image captioning. *arXiv preprint arXiv:2210.04559*.

Xu Yan, Zhengcong Fei, Zekang Li, Shuhui Wang, Qingming Huang, and Qi Tian. 2021. Semi-autoregressive image captioning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2708–2716.

Xu Yang, Hanwang Zhang, and Jianfei Cai. 2019. Learning to collocate neural modules for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4250–4260.

Hong Yu, Yuanqiu Liu, Baokun Qi, Zhaolong Hu, and Han Liu. 2023. End-to-end non-autoregressive image captioning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588.

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 *Proceedings of the AAAI conference on artificial intelligence*, pages 13041–13049.

Yuanen Zhou, Yong Zhang, Zhenzhen Hu, and Meng Wang. 2021. Semi-autoregressive transformer for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3139–3143.

Zixin Zhu, Yixuan Wei, Jianfeng Wang, Zhe Gan, Zheng Zhang, Le Wang, Gang Hua, Lijuan Wang, Zicheng Liu, and Han Hu. 2022. Exploring discrete diffusion models for image captioning. *arXiv preprint arXiv:2211.11694*.

A Supplementary materials

A.1 Code

Our code is uploaded via an anonymous link: "https://anonymous.4open.science/r/NAR-IC-ARR24".

A.2 Preliminary for AR and NAR

The conditional probabilities of the generated caption $\boldsymbol{Y}$ are defined as:

$$p(\boldsymbol{Y}|I;\theta) = \begin{cases} \prod_{i=1}^{|\boldsymbol{Y}|} p\left(y_i|y_1,...,y_{i-1},I;\theta\right), & AR; \\ \prod_{i=1}^{|\boldsymbol{Y}|} p\left(y_i|I;\theta\right), & NAR. \end{cases} \tag{6}$$

AR models generate the subsequent word y_i based on the previous context $y_1, ..., y_{i-1}$. It determines that the inference process is not parallelizable. Unlike AR, NAR eliminates sequential dependencies, and the generated sentence depends solely on the image. When y_i and $y_1 : y_{i-1}$ are independent, the conditional probabilities are degenerated and this inference process is parallelizable.

A.3 Experimental settings

Consistent with the most IC models, we convert all the captions to lowercase and remove words that occur fewer than 6 times. The remaining 9487 words constitute our vocabulary. We set the maximum sentence length L to 16, the embedding size of the model d to 512, the number of the encoder and decoder layers N_E and N_D to 3, and the number of Transformer heads h to 8. The image feature ($n \times n$, feat_dim) is extracted by the pre-trained ViT/Swin Transformer, shaped as $(16 \times 16, 1024)/(12 \times 12, 1536)$. We employ the Adam optimizer (Kingma and Ba, 2014) with a warm-up period of 10,000 iterations. The batch size is set to 256, and the learning rate is initialized at 5×10^{-3}. The learning rate undergoes decay by a factor of 0.8 every 3 epochs.

The total training epochs are set to 200 under cross-entropy loss. It is trained on 4 NVIDIA V100 GPUs, and the whole training process takes about 80 GPU hours. Here we provide more details about the settings about the NAR+AR, AR+NAR, and NAR+Mixed in Table 1. In the NAR+AR mode, we train NAR for 100 epochs first, followed by training AR for another 100 epochs. Conversely, in the AR+NAR mode, we train AR for 100 epochs initially, followed by NAR for another 100 epochs. In the NAR+Mixed approach, we train NAR for 100 epochs initially. Subsequently, we alternate between training AR and NAR for 10 epochs each until the total epoch count reaches 200. Additional 20 epochs for RL training is applied for fair comparison (only used in Table 2 and Table 3). The "Reduce-On-Plateau" strategy is applied with a decay rate of 0.5 and patience of 3.

A.4 Qualitative Results

Besides, we provide some examples of the CIDEr (Vedantam et al., 2015), ROUGE (ROUGE, 2004), and CLIP (Xu, 2022) scores for label selection from MSCOCO (Lin et al., 2014) and knowledge distillation, as illustrated in Figure 4. Upon observing these examples, it is evident that the results of CIDEr and ROUGE selection are generally consistent. In Figure 4 (a), we choose the third annotation through CIDEr and ROUGE, while opting for the fourth annotation based on its CLIP score. While in Figure 4 (b), the first/fifth/forth annotations are selected by CIDEr/ROUGE/CLIP score,respectively. Figures 4 (c) and (d) present a situation where the AR-IC model predicts the ground-truth labels correctly. Although the label predicted by the AR-IC model is not entirely identical to the ground-truth in Figures 4 (e), they have almost the same sentence structure and content. Furthermore, annotations with higher CIDEr and ROUGE scores tend to exhibit a mid-to-high CLIP score. This observation also provides additional verification that some annotations from MSCOCO may have certain sentence pattern that are not well-suited for NAR-IC training. For example, the first, second, and fifth annotations in Figure 4 (a) are deemed poor under all three evaluation metrics.

Figure 4: Examples of the CIDEr, ROUGE, and CLIP score in label selection and knowledge distillation.