Image Captioning With Controllable and Adaptive Length Levels

Ning Ding, Chaorui Deng, Mingkui Tan, Qing Du, Zhiwei Ge, and Qi Wu

***Abstract*—Image captioning is a core challenge in computer vision, attracting significant attention. Traditional methods prioritize caption quality, often overlooking style control. Our research enhances method controllability, enabling descriptions of varying detail. By integrating a length level embedding into current models, they can produce detailed or concise captions, increasing diversity. We introduce a length-level reranking transformer to correlate image and text complexity, optimizing caption length for informativeness without redundancy. Additionally, with caption length increase, computational complexity grows due to the autoregressive (AR) design of existing methods. To address this, our non-autoregressive (NAR) model maintains constant complexity regardless of caption length. We've developed a training approach that includes refinement sequence training and sequence-level knowledge distillation to close the performance gap between NAR and AR models. In testing, our models set new standards for caption quality on the MS COCO dataset and offer enhanced controllability and diversity. Our NAR model excels over AR models in these aspects and shows greater efficiency with longer captions. With advanced training techniques, our NAR's caption quality rivals that of leading AR models.**

***Index Terms*—Length-controllable image captioning, non-autoregressive image captioning, length level reranking, refinement-enhanced sequence training.**

I. Introduction

THE task of Image captioning is to automatically describe a given image with a natural sentence, and has developed rapidly thanks to the remarkable progress in deep learning methods and open datasets. It is a challenging task as it requires a comprehensive understanding of the image content as well as a strong ability of natural language expression. It is also an important task in practice and has wide applications such as text-to-image retrieval, multi-modal recommendation, and human-computer interaction, *etc*. State-of-the-art (SOTA) methods in image captioning prone to the Encoder-Decoder framework [1], [2], [3], [4], [5], [6], where an encoder extracts features from the input image, followed by a decoder that generates captions based on the encoder features in a *autoregressive* manner, i.e., predicting one token at each step. Based on this framework, remarkable performances have been achieved on the challenging MS COCO dataset [7], and even surpass human performance on some evaluation metrics.

Despite this, most of these SOTA methods lack the ability to control the style of the generated image captions; more especially, choosing to caption the image at a specified level of detail. As shown in Fig. 1(a), given an input image, although the caption generated by VLP [6] (a current SOTA) correctly describes the image, it fails to capture more informative visual concepts such as "pitcher throws a pitch" and "wearing a uniform", which also leads to a limited diversity. This motivates us to develop controllable image captioning models that can generate as requested either rough or detailed image captions. We show

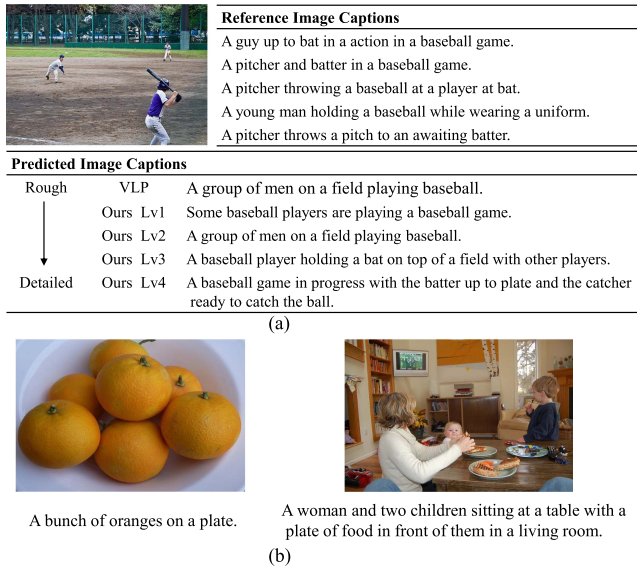


Fig. 1. (a) Illustration of image captions with different lengths. To the right of the image are five human-annotated captions. At the bottom, we show the image captions generated by an original VLP [6] model and our length-controllable version of VLP. (b) Illustration of the semantic complexity correlation: semantically simple images can be described with short & brief captions; on the contrary, long & detailed captions are required for more complex images.

Manuscript received 30 June 2022; revised 20 September 2023; accepted 11 October 2023. Date of publication 6 November 2023; date of current version 8 January 2024. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62072190, in part by the Key-Area Research and Development Program of Guangdong Province under Grants 2018B010107001 and 2019B010155001, in part by the Ministry of Science and Technology Foundation Project under Grant 2020AAA0106900, and in part by the Program for Guangdong Introducing Innovative and Enterpreneurial Teams under Grant 2017ZT07X183. Recommended for acceptance by L. Wang. *(Ning Ding and Chaorui Deng contributed equally to this paper.) (Corresponding author: Mingkui Tan.)*

Ning Ding and Zhiwei Ge are with JD.com, Beijing 101111, China (e-mail: dingning36@jd.com; gezhiwei@jd.com).

Chaorui Deng and Qi Wu are with the School of Computer Science, The University of Adelaide, Adelaide, SA 5005, Australia (e-mail: chaorui.deng@adelaide.edu.au; qi.wu01@adelaide.edu.au).

Mingkui Tan and Qing Du are with the School of Software Engineering, South China University of Technology, Guangzhou 510006, China (e-mail: mingkuitan@scut.edu.cn; duqing@scut.edu.cn).

Digital Object Identifier 10.1109/TPAMI.2023.3328298

in this paper that such an ability can be effectively acquired by directly controlling the length of the generated image captions.

Length is an important property of natural language since it roughly reflects the amount of information carried by a sentence. In this work, we explicitly exploit this property and propose a length-controllable image captioning approach that can be applied to existing image captioning models seamlessly. See the example in Fig. 1(a), the longest caption (Ours Lv4) generated by our length-controllable VLP contains detailed descriptions of the salient objects, leading to higher fidelity of the visual information. While short captions (Ours Lv1/Lv2) briefly introduce the scene but can be generated more efficiently. Besides, we further design a reranking module that is able to find the most suitable level of detail for each image according to its semantic complexity. As shown in Fig. 1(b), when the image is semantically simple, a short caption may properly describe the image, while a long caption would be torturous. On the contrary, for a complex image, a long caption can capture the visual semantics more comprehensively, while a short one will inevitably lose some important details.

At the core of our method is a concept termed "*length level*" which refers to a specific length range of the image captions. Specifically, during training, a length-level embedding is trained for each level with only the training data inside the length range. Thus, the model is enabled to capture the language patterns of the captions on each level, e.g., longer captions tend to involve more visual concepts. During inference, based on different length level embeddings, the model is controlled to generate image captions within different length ranges. In this way, an existing image captioning model can be turned into a length-controllable one by simply introducing an additional length level embedding to the input. Afterward, a natural extension is to select a proper length level when describing the image so that the image caption is informative while also not redundant. We achieve this by proposing a length-level reranking transformer (LLRT), which takes as input the image as well as the generated captions at all length levels and predicts the most suitable length level. A special `[LEVEL]` token is appended to the input of LLRT, and its final hidden state is fed into a scoring head to score the caption.

We show the effectiveness of our length-controllable and length-level reranking approaches by applying them to several popular image captioning models under both Teacher Forcing training (AoANet [4] and VLP [6]) and Self-Critical Sequence Training (SCST) [8] schemes (M^2 Transformer [9] and X-LAN [10]). In the experiments, our length-controllable models successfully generate high-quality and length-controllable results. By further reranking the captions from all length levels, we obtain top-1 performances that are significantly higher than the original results of these baseline models. Nevertheless, a new problem appears for these models: since they adopt an *autoregressive* (AR) decoding strategy, which generates only one token at each step, their decoding complexity increases linearly as the length L of the caption grows (i.e., a $\Theta(L)$ complexity). This hampers the model efficiency in scenarios where longer captions are preferred, and also impedes the parallel generation of image captions with different length levels.

To tackle this, we propose a *non-autoregressive* (NAR) paradigm for length-controllable image captioning, denoted by LaNAR (Length-aware Non-Autoregressive) Captioning. Specifically, the proposed LaNAR paradigm decodes image captions within a fixed number of refined steps regardless of L, which is a *length-irrelevant* complexity. Moreover, LaNAR is compatible with transformer-based architectures as they have the potential to process the whole input sequence in parallel. We verify LaNAR on two architectures, including the encoder-decoder-based vanilla Transformer [11], and the pure encoder-based BERT [12]-like architecture. From the experiments, the proposed LaNAR paradigm significantly improves the decoding efficiency for longer captions, while also achieving competitive or even better performance compared with the AR baselines on all length levels. We further devise a refinement-enhanced sequence training (REST) scheme specially for our LaNAR paradigm, which significantly improves its performance and outperforms existing NAR image captioning models by a large margin. After applying REST and Sequence-Level Knowledge Distillation (SLKD) [13], our LaNAR captioning models outperform their AR counterparts while decoding much faster.

Our main contributions are summarized as follows:

1) We first introduce the design of "length level" as a control signal to learn length-aware image captioning models, which can be easily integrated into existing image captioning methods to make them capable of generating high-quality, length-controllable and diverse image captions.
2) We devise LaNAR, a NAR paradigm for length-controllable image captioning that makes the decoding of long captions more efficient while also achieving higher control precision and producing more diverse results than the AR baselines.
3) We propose to learn a length level reranking transformer to find out the most suitable length level when depicting a given image.
4) We perform extensive experiments on various kinds of image captioning models and settings to show the effectiveness of our proposed methods.

A preliminary version of this work was published in ECCV2020 [14]. This work is a systematic extension of our previous paper in the following ways:

1) We devise a length-level reranking transformer that is able to caption the image within a proper length level according to its semantic complexity.
2) We devise the REST scheme and adopt the SLKD technique for training LaNAR captioning models, which bridges the gap between the NAR model and AR model, boosting the performance of LaNAR captioning models by a large margin.
3) More experimental results are provided, including the performance of the proposed length-controllable paradigm on more baseline models, the results of LLRT on both length-controllable AR and NAR models, the results of the length-controllable AR models under SCST, and the results of REST and SLKD on NAR models, as well as more performance discussions.

II. Related Works

A. AR-Based Image Captioning

Over the years, auto-regressive (AR) Image Captioning methods have developed rapidly based on the Encoder-Decoder framework [1], [2]. In [15], the authors proposed to integrate attention mechanisms [16] into RNN-based decoders, to encourage the model to focus on special parts of the image during each decoding step. Then, Anderson et al. [17] devised a bottom-up attention mechanism to enable feature aggregation at object-level, instead of pixel-level as in [15]. This method achieved the best results at that time and outperformed the second-best result by a large margin. In [8], Rennie et al. developed a Reinforcement Learning based training strategy that directly optimizes the CIDEr [18] score of the predicted image captions through policy gradient. The proposed method, called self-critical sequence training (SCST), greatly alleviates the "exposure bias" problem in sequence modeling and significantly boosts performance. Some works, on the other hand, focus on leveraging additional information, such as semantic attributes [19] and visual relations [20], [21], to improve the image caption quality. More recently, after witnessing the effectiveness of Transformers [11] in capturing long-range dependencies in sequence modeling, many Transformer-based methods [4], [5], [6] have been proposed to further advance the image captioning performance.

B. Diverse and Controllable Image Captioning

Despite existing AR image captioning models having achieved remarkable performance, fewer efforts have been made toward improving the diversity and controllability of image captions. In [22], a Part-of-Speech (POS) predictor is trained to generate a sequence of POS tags based on the input images, which are then used to control the decoding of image captions. Chen et al. [23] proposed to control the image captions through the Abstract Scene Graph (ASG), which is a directed graph consisting of three types of abstract nodes (object, attribute, relationship) grounded in the image. However, these methods rely on additional tools or annotations to provide supervision, and their control signals are too abstract to be used conveniently in practice. Some content-based methods [24], [25], [26] used different image regions to generate region-specific image captions, but they lack the ability of fine-grained control. Besides, Generative Adversarial Network (GAN) based methods have also been proposed to generate image captions with diverse styles [27], [28], [29]. These methods require additional training data, such as an image caption dataset with additional style annotations [30], [31], [32], which is scarce and expensive; or a large corpus of stylized text without aligned images [33], [34], which often leads to unsatisfied caption quality.

As discussed in Section I, length is an important property for image captions. It is easy to acquire and is strongly associated with the semantic complexity of the image caption, which is very useful in practice. Several methods in the Natural Language Processing field have visited the length-controllable text generation setting. In [35], Kikuchi et al. explored four length-control strategies in Neural Sentence Summarization: 1) omit the `[EOS]` token during decoding until the desired length is reached; 2) set a length range and manually discard out-of-range sequences; 3) use a *remained length* embedding to inform the model of the remained length; 4) multiply the length with the hidden state during parameter initialization. However, in the first two strategies, the model is not aware of the desired length, which may lead to low diversity. Also, these two strategies are likely to produce uncompleted sentences. The last two strategies seek to control the exact length of the output sentence, which is hard in practice and restricts the flexibility of the results. [36] also controls the exact length of the output in Convolutional Seq2Seq models [37], which faces the similar problems.

Different from the above methods, we propose to use the *length level* as a control signal to obtain diverse and controllable image captions, which is more convenient compared with those using abstract control signals like POS tags and ASG, cheaper compared with those requiring additional annotations and tools, and more applicable compared with those controlling the exact length of the sentence.

C. NAR Text Generation

A common problem in AR models is that the output tokens must be generated sequentially, which prevents architectures like the Transformer from fully realizing their training-time speedup advantage during inference. To tackle this, some recent Neural Machine Translation works have appealed to NAR decoding algorithms [38], [39], [40], [41], [42], which attempts to predict the entire sequence within one (or, a fixed number irrelevant to the sequence length) forward pass of the decoder. Unlike AR models where the decoding process terminates automatically after encountering the `[EOS]` token, existing NAR models usually need to determine the length of the output sequence at the beginning of the decoding process. These methods either learn a length predictor along with the decoder or adopt insertion/deletion modules to automatically change the length of the output. More recently, several NAR image captioning models have been developed [43], [44], however, they usually suffer from a large performance degradation compared with the SOTA AR image captioning models.

In this paper, we devise a NAR approach for length-controllable image captioning to reduce the decoding complexity, especially for long captions. Thanks to the design of "length level" and the explicit modeling of the length property, our approach is able to automatically find a suitable end position within the length level during decoding, without having to learn an additional length predictor. Moreover, we propose a refinement-enhanced sequence training scheme for our NAR image captioning model, where we achieve competitive performance with SOTA AR models.

III. Method

A. Preliminary

Given an image $\boldsymbol{I}$, the target of image captioning is to generate a natural sentence description $\boldsymbol{S} = \{s_i\}_{i=1}^{L}$ for $\boldsymbol{I}$. Here, s_i is a token in $\boldsymbol{S}$, L is the length of $\boldsymbol{S}$. Existing methods are mostly

autoregressive, where they factorize the distribution of $\boldsymbol{S}$ into a chain of conditional probabilities with a left-to-right causal structure: $p(\boldsymbol{S}|\boldsymbol{I}) = \prod_{i=1}^{L} p(s_i|s_{j<i}, \boldsymbol{I})$. Consequently, a token s_i can only be generated when all preceding tokens $s_{j<i}$ are available. Assume the target image caption to be $\boldsymbol{S}^* = \{s_i^*\}_{i=1}^{L^*}$. The training of AR models typically follows the "Teacher Forcing" [45] scheme, which aims to minimize the negative log-likelihood of the ground-truth token s_i^* given all preceding ground-truth tokens $s_{j<i}^*$

$$\min \sum_{i=1}^{L^*} -\log p(s_i^*|s_{j<i}^*, \boldsymbol{I}). \tag{1}$$

Apart from Teacher Forcing, a sequence-level optimization scheme SCST is proposed in [8], which seeks to maximize the expected reward of the predicted sentences

$$\max E_{\boldsymbol{S}\sim\pi}[r(\tilde{\boldsymbol{S}})], \tag{2}$$

where π indicates the output distribution modeled by the image captioning model, and the reward function $r(\cdot)$ is defined as the subtraction of the CIDEr score of the randomly sampled $\tilde{\boldsymbol{S}}$ and their greedy sampled counterpart $\hat{\boldsymbol{S}}$, i.e., a self-critical reward

$$r(\tilde{\boldsymbol{S}}) = \text{CIDEr}(\tilde{\boldsymbol{S}}) - \text{CIDEr}(\hat{\boldsymbol{S}}). \tag{3}$$

In datasets like MS COCO, each training image usually has 5 paired ground-truth captions, and the CIDEr score of a generated caption is calculated over all of them.

During inference, AR models start by taking a special `[BOS]` token as input and predict the sentence one token after another, until a special `[EOS]` token is reached.

B. Acquisition of Length Information in AR Image Captioning

To explicitly model the length property of an image caption $\boldsymbol{S}$, we assign $\boldsymbol{S} = \{s_i\}_{i=1}^{L}$ into a specific length level k according to its length L, which has a length range of $[L_k^{low}, L_k^{high}]$. Each length level is associated with a length level embedding $\boldsymbol{e}_l(k) \in \mathbb{R}^d$ to differentiate image captions on different length levels. Then, for each token s_i in $\boldsymbol{S}$, we construct its input embedding by

$$\boldsymbol{x}_{s_i} = \boldsymbol{e}_l(k) + \boldsymbol{e}_w(s_i) + \boldsymbol{e}_p(i), \tag{4}$$

where $\boldsymbol{e}_w(s_i) \in \mathbb{R}^d$ is the word embedding of s_i and $\boldsymbol{e}_p(i) \in \mathbb{R}^d$ is an optional positional embedding for Transformer-based decoder. With this length level embedding $\boldsymbol{e}_l(k)$, the length information of $\boldsymbol{S}$ is explicitly incorporated into $\boldsymbol{x}_{s_i}$. In this way, existing image captioning models can be turned into length-aware models by simply replacing their original token embeddings (e.g., word embeddings) with our length-aware tokens embeddings, without any other modifications to their network architectures.

The training of length-aware AR image captioning models can directly follow the Teacher Forcing or SCST scheme. During training, the length level embedding for level k will only be trained with captions within the particular length range $[L_k^{low}, L_k^{high}]$. For Teacher Forcing, the only difference is to separate the training set according to the length levels; similarly, for SCST, the CIDEr score of a generated caption will only be calculated with the reference captions in the same length level. In this way, the "trait" of image captions with different length levels is separately modeled, e.g., long captions usually cover more visual concepts in the image, which enables length-aware vision-language modeling. Note that, each image is only trained on the length levels that have at least one reference caption. During inference, the desired length level embedding is fed into the model as a control signal, through (4).

Due to the simplicity of the proposed length-controllable approach, it can be easily implemented on existing AR image captioning methods. To show its strong generalization ability, we first consider two representative baseline models, i.e., an LSTM-based AoANet [4] and a Transformer-based VLP [6], and train these models using the Teacher Forcing scheme; moreover, we further investigate another two models, i.e., M2 Transformer [9] and X-LAN [10], under the self-critical training scheme. As an example, we illustrate the length-controllable VLP in Fig. 2(a). When setting the boundary $[L_k^{low}, L_k^{high}]$ of a length level we follow two simple principles: 1) there should be enough training data for each length level so as to train the length level embedding sufficiently; 2) the range of a length level should not be too narrow to ensure the flexibility of the generated captions. After checking the length distribution of captions in the MS COCO dataset (see Fig. 5), we explore two length-level division plans in our experiments, which contain 4 or 5 length levels, respectively. The 4-level plan divides the image captions into 4 chunks with length inside the ranges `[1,9]`, `[10,14]`, `[15,19]`, and `[20,25]`, respectively, from rough to detailed. While the 5-level plan provides more fine-grained divisions, which are `[1,9]`, `[10,13]`, `[14,17]`, `[18,21]` and `[22,25]`.

C. Adaptive Length Level Reranking

So far, we have introduced our length-controllable image captioning approach, which requires a "length level" control signal to decide the length range of generated captions. However, as shown in Fig. 1(b), for a semantically simple image, a short sentence is usually enough to cover all the details and a long sentence could be tedious and unnecessary, while for a complex image, short sentences may only able to describe the image from a coarse global view and fail to capture the distinct part of the image. Thus, it is intuitive to capture the correlation between the semantic complexities of the image modality and the text modality, and adaptively select the most suitable length level for an image. While there have been multiple reference-free automatic image captioning metrics that can be used to rank the image captions, such as CLIPScore [46], VIFIDEL [47], and UMIC [48]. However, they focus on the *semantic alignment* between the generated captions and the images, which may fail to capture the subtle difference in terms of the *semantic complexities*. More critically, the measurement of the semantic complexity of an image is still an open problem.

To tackle this, we take inspiration from [49], which shows that the complexity of a thing is correlated to how hard it is for a human to describe it. Therefore, the most suitable length level of an image can be determined by ranking the corresponding captions according to a reference-based evaluation metric with

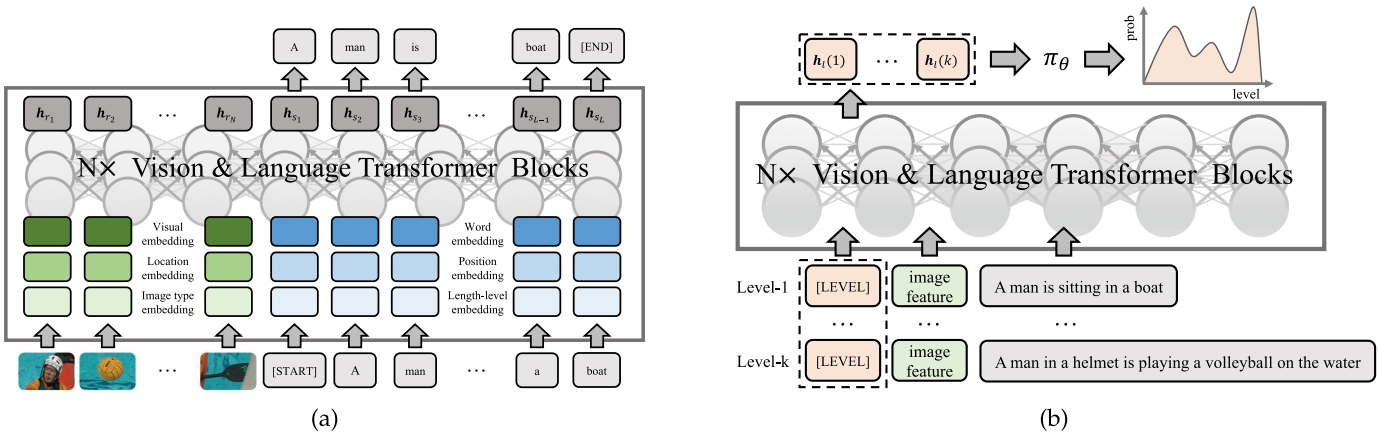


Fig. 2. (a) Overview of length-controllable VLP. The caption tokens are represented by the sum of the word embeddings, positional embeddings, and the proposed length-level embeddings. $\boldsymbol{h}_r$ and $\boldsymbol{h}_s$ are the output hidden states of image tokens and caption tokens. (b) Overview of the length level reranking transformer (LLRT). The generated captions from all length levels are fed into LLRT simultaneously, and a length level distribution is predicted from their [LEVEL] token embeddings through a scoring head π_θ.

human-provided references. To this end, we learn a Length-level Reranking Transformer (LLRT) module with CIDEr optimization. As shown in Fig. 2(b), LLRT adopts a joint vision-language transformer architecture, which takes as input the image $\boldsymbol{I}$, the generated caption of length level k (denoted by $\boldsymbol{S}_k$), as well as a special [LEVEL] token. Based on the final hidden state of the [LEVEL] token, denoted by $\boldsymbol{h}_l(k) \in \mathbb{R}^d$, we then predict a confidence score between $\boldsymbol{S}_k$ and $\boldsymbol{I}$ from $\boldsymbol{h}_k$ through a scoring head, and choose the $\boldsymbol{S}_k$ with the largest score as the final prediction for $\boldsymbol{I}$. Note that, since the original VLP model is already a joint vision-language transformer, we can directly reuse the backbone of VLP and attach the scoring head on top of it.

Specifically, we formulate the optimal length level selection as a reinforcement learning problem. Given the generated image captions from all length levels $\{\boldsymbol{S}_k\}_{k=1}^K$ (K is the number of length levels), we first obtain their [LEVEL] representations $\boldsymbol{H}_l = \{\boldsymbol{h}_l(k)\}_{k=1}^K$ through the LLRT backbone, and compute the CIDEr scores between each $\boldsymbol{S}_k$ and the human-provided reference captions, denoted by $\text{CIDEr}(\boldsymbol{S}_k)$. Taking $\text{CIDEr}(\boldsymbol{S}_k)$ as reward, our learning objective is to minimize the negative expected reward

$$L_r = -\mathbb{E}_{k\sim\pi_\theta(\boldsymbol{H}_l)}[\text{CIDEr}(\boldsymbol{S}_k)], \tag{5}$$

where $\pi_\theta(\cdot)$ is a two-layer Multi-Layer Perceptron with parameter θ that takes $\boldsymbol{H}_l$ as input and output a confidence score between $\boldsymbol{S}_k$ and $\boldsymbol{I}$ for $k = 1, \ldots, K$. Afterward, a Softmax function is adopted to compute the categorical length level distribution based on the confidence scores of all length levels. The model can be optimized through policy gradient [50] as follows:

$$\nabla_\theta L_r = -\text{CIDEr}(\boldsymbol{S}_k)\nabla \log \pi_\theta(\boldsymbol{H}_l). \tag{6}$$

When training LLRT, we can initialize it from a trained length-controllable VLP model, and fine-tune it with (5) while keeping the original parameters in length-controllable VLP fixed and only training the scoring MLP head. The token embeddings are also fixed during fine-tuning except for the newly added [LEVEL] token. The [LEVEL] token embedding is initialized from the original [CLS] token embedding in length-controllable VLP. During inference, the reranking process requires only one forward pass of the LLRT model, which is very efficient compared with the caption decoding process.

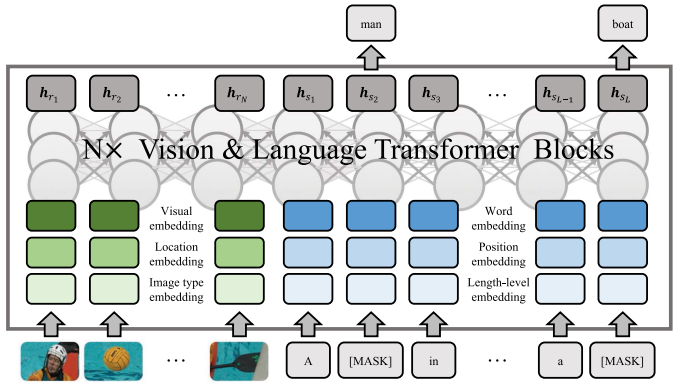


Fig. 3. Overview of LaNAR-VLP. The network architecture is similar to the length-controllable VLP, but the input is randomly masked and the final hidden state of [MASK] tokens are fed into a token classifier to predict their original tokens.

D. NAR Length-Controllable Image Captioning

AR image captioning models suffer from a linearly-increased decoding complexity which usually leads to inefficiency, especially for long caption generation. To aid this, we propose a NAR length-controllable image captioning paradigm named LaNAR for transformer-based image caption models. Take the BERT [12] model as an example, the architecture of the LaNAR-BERT is shown in Fig. 3. Following [6], [17], the input image $\boldsymbol{I}$ is first pre-processed by a pre-trained object detector into M object proposals $\boldsymbol{R} = \{r_i\}_{i=1}^M$. These proposals are represented by their region features $\boldsymbol{F}_e = \{\boldsymbol{f}_e(i)\}_{i=1}^M$, classification probabilities $\boldsymbol{F}_c = \{\boldsymbol{f}_c(i)\}_{i=1}^M$, and localization features $\boldsymbol{F}_p = \{\boldsymbol{f}_p(i)\}_{i=1}^M$. Similar as in [6], the final input representation of

r_i is constructed by

$$\boldsymbol{x}_{r_i} = \boldsymbol{W}_e^T \boldsymbol{f}_e(i) + \boldsymbol{W}_p^T[\text{LN}(\boldsymbol{f}_c(i)), \text{LN}(\boldsymbol{f}_p(i))] + \boldsymbol{e}_{img}. \tag{7}$$

$[\cdot,\cdot]$ indicates the concatenate operation, and LN represents Layer Normalization [51]. $\boldsymbol{W}_e$ and $\boldsymbol{W}_p$ are two learnable projection matrices that project the corresponding features into d-D visual embeddings and location embeddings, respectively. $\boldsymbol{e}_{img} \in \mathbb{R}^d$ is a learnable embedding that differentiates the image regions from the text tokens, which plays a similar role as the segment embeddings in BERT. We also apply the LaNAR paradigm to the vanilla Transformer [11]. It is an encoder-decoder-based architecture, where the encoder is used to process the image features with a similar process as in (7) (except $\boldsymbol{e}_{img}$ is removed), and similar modifications as in (4) is applied to the decoder part. We denote this model as LaNAR-Transformer.

Training. The training of LaNAR captioning models follows the basic idea of Conditional Masked Language Modeling [39], but we make modifications to take advantage of our length level design and get rid of the commonly used length predictor. Given the target image caption $\boldsymbol{S}^*$ inside the length level $[L^{low}, L^{high}]$, we first pad it with `[EOS]` tokens to the longest length L^{high}. Then, we randomly choose $m \in [1, L^{high}]$ positions in the sequence and replace them with the `[MASK]` token. Denoting the obtained sequence as $\boldsymbol{S}^m$, the LaNAR captioning model attempts to predict the original tokens at all masked positions conditioned only on the image region representations (obtained by (7)) and the unmasked length-aware tokens embeddings in $\boldsymbol{S}^m$ (obtained by (4)). Hence, the predicted conditional probabilities are independent of each other, allowing them to be calculated in parallel at inference time. We train the LaNAR captioning model by minimizing the cross-entropy loss over all masked positions

$$L_c = \sum_{i=1}^{L^{high}} -\mathbb{1}(s_i) \cdot \Omega(s_i^*) \cdot \log p(s_i = s_i^*). \tag{8}$$

Here, $\mathbb{1}(\cdot)$ is an indicator function that outputs 1 if $s_i =$ `[MASK]` and 0 otherwise. To facilitate the model to produce longer captions, we adopt a term $\Omega(s_i^*)$ in (8) that outputs ω ($\omega < 1$) when s_i^* is `[EOS]` and 1 otherwise, so that the gradient contributed by `[EOS]` tokens is down-scaled, making the model less likely to predict the `[EOS]` token.

By padding $\boldsymbol{S}^*$ to L^{high} with `[EOS]` tokens, the proposed LaNAR paradigm is trained to automatically find a suitable end position within $[L^{low}, L^{high}]$, since all training samples for this length level only contain the `[EOS]` tokens inside $[L^{low}, L^{high}]$. Owing to this design, a LaNAR captioning model does not need a length predictor to determine the length of the output at the start of decoding as most NAR text generation methods do, but also shows clearly better controllability compared with our length-controllable AR baselines (Section IV-H).

Inference. We perform parallel image caption decoding based on the idea of iterative refinement [39], [43], whereat each step t, a masked image caption $\boldsymbol{S}_{t-1}^m$ (obtained from the previous step) is fed into the model to predict the tokens in the masked positions. Specifically, the LaNAR captioning model predicts the token distribution for all positions in $\boldsymbol{S}_{t-1}^m$, denoted by p_t. Then, we update all masked positions in $\boldsymbol{S}_{t-1}^m$ with the greedy-sampled predicted tokens

$$s_i \leftarrow \arg\max_s p_t(s_i = s), \ \forall i \in \{i | s_i = \texttt{[MASK]}\}. \tag{9}$$

Denote the updated caption as $\boldsymbol{S}_t$. To encourage the model to predict longer captions, before the greedy sampling, we exponentially decay the probability of the `[EOS]` token by a factor γ for positions after L^{low}

$$p_t(s_i = \texttt{[EOS]}) \leftarrow \gamma^{L^{high}-i} p_t(s_i = \texttt{[EOS]}), \quad \forall i \in [L^{low}, L^{high}]. \tag{10}$$

Meanwhile, based on p_t, we also obtain a confidence score $c_{t,i}$ for each token s_i

$$c_{t,i} \leftarrow \begin{cases} \max_s p_t(s_i = s), & i \text{ is a masked position.} \\ (1-\alpha) * c_{t-1,i} + \alpha * \max_s p_t(s_i = s), & \text{otherwise.} \end{cases} \tag{11}$$

We then find the tokens with the lowest n confidence scores and mask the corresponding positions, resulting in $\boldsymbol{S}_t^m$, which will be fed into the LaNAR captioning model again for next-step refinement. Here, α is a hyper-parameter that controls the progressing speed of the confidence scores in unmasked positions. Let T be the overall refine steps, $n = \frac{T-t}{T} L^{high}$ is the number of masked positions and will decay linearly to 0 as t increases. The initial image caption $\boldsymbol{S}_0^m$ is set as $\{s_i = \texttt{[MASK]}\}_{i=1}^{L^{high}}$. An illustration of the iterative refinement process is in Fig. 4.

Through iterative refinement, the decoding complexity is decreased from $\Theta(L^{high})$ in AR methods to $\Theta(T)$ in LaNAR models. Also, the mistakes made at early steps in LaNAR models are possible to be revised in future steps, which is infeasible for AR methods. Note that the update rule in (11) is different from the update rule in [39], which only updates the confidence scores of the masked positions. In practice, we found ours (denoted as the *global update rule*) performs much better in terms of caption quality. Moreover, our LaNAR paradigm also allows dynamic length changes during the refinement process, while not using any additional insertion/deletion modules like in [40].

E. Self-Critical Training for Iterative Refinement

Sequence-level optimization like SCST has been shown to greatly improve the performance of AR models. However, it is non-trivial to apply SCST directly to the LaNAR paradigm for its non-autoregressive decoding behavior. To aid this, we design a Refinement-Enhanced Sequence Training scheme for LaNAR, denoted by REST. Specifically, at each refine step T, given the masked caption obtained at the previous step $\boldsymbol{S}_{t-1}^m$, we obtain two new captions by one forward pass of the LaNAR captioning model, i.e., $\tilde{\boldsymbol{S}}_t$, where the updated token for each masked position in $\boldsymbol{S}_{t-1}^m$ is randomly sampled from p_t, and $\hat{\boldsymbol{S}}_t$, where the tokens are greedy-sampled. Then, we follow SCST to maximize the expected reward of the updated sentence as in (2), where the reward function is defined as

$$r(\tilde{\boldsymbol{S}}_t) = (\text{CIDEr}(\tilde{\boldsymbol{S}}_t) - \text{CIDEr}(\hat{\boldsymbol{S}}_t)) + (\text{CIDEr}(\tilde{\boldsymbol{S}}_t) - \text{CIDEr}(\hat{\boldsymbol{S}}_{t-1})). \tag{12}$$

Mask-Predict-Update Process:

Step 1	[MASK] [MASK][MASK][MASK] [MASK] [MASK][MASK]…[MASK] [MASK][MASK]
	Three children of at sitting at a table table pizza pizza pizza pizza pizza a a slice of pizza.
	Three children sitting at a table with pizza pizza pizza pizza pizza and and a a slice of them.
	Three children sitting at a table with a slice of pizza and a paper plate in front of them.
Step T	Three children sitting at a table with a slice of pizza and a plastic cup in front of them.

Fig. 4. Illustration of the iterative refinement process of our LaNAR paradigm. The red tokens indicate those with lower confidence scores. At each step, all red tokens are masked and re-predicted in parallel, conditioned on other tokens in the sequence and visual information from the image.

The first term is inherited from the original SCST, where we compute the advantage of the randomly sampled $\tilde{\boldsymbol{S}}_t$ over its greedy sampled counterpart $\hat{\boldsymbol{S}}_t$; and the second term is our refinement-enhanced reward, where the advantage of $\tilde{\boldsymbol{S}}_t$ over the caption generated at the previous step is considered. This encourages the model to produce better captions after each refine step.

During the REST procedure, a LaNAR captioning model takes as input an image, and the length level index k of a randomly sampled reference caption of the image. Based on k, a masked sequence of length L_k^{high} is initialized as $\boldsymbol{S}_0$, and the CIDEr score is only computed with the reference captions within the same length level, similar as in the self-critical training of length-controllable AR models. We refine each sentence 10 times during REST, where each step recovers 10% of the masked tokens. Since $\boldsymbol{S}_0$ is a sequence of `[MASK]` tokens, the proposed REST is not used for the first refine step.

F. Sequence-Level Knowledge Distillation

NAR text generation models usually suffer from performance degradation compared with AR models, since the tokens are generated (semi-)independently and thus their sequential dependencies are not as well-captured as in AR models. The key reason for this problem is the "multi-modality" of the training data, i.e., a source image can be captioned in various ways. As shown in [38], [52], Sequence-Level Knowledge Distillation (SLKD) [13] is an effective strategy to reduce the modes of the training corpus and alleviate the multi-modality problem. Essentially, these methods use the predicted sentences from the AR teacher model to construct a new dataset to supervise the training process of the NAR student model. The new dataset provides less noisy and more deterministic image captions which make the NAR student model easier to learn.

In the case of our length-controllable NAR image captioning, the simplest way to apply SLKD is to use an existing AR image captioning model as the teacher and generate a new and mode-reduced training set. However, the captions on each length level usually have specific modes, while existing AR models can only produce data with modes concentrated on the dominant length level (see Fig. 8(b)), which is inappropriate for our setting. Therefore, we propose to use a length-controllable AR model, i.e., a length-controllable VLP as the teacher. Specifically, we adopt the teacher model to generate multiple captions with different length levels for each training image, and randomly subsample the generated captions following the length distribution of the original training set. We show in the experiments (Table IV) that it is important to maintain the length distribution of the original training set for a good performance. Moreover, since the subsampling process largely reduces the amount of the new training data, we merge the subsampled data with the original training set as the new training set, which is different from [38], [52] where they merely use the newly generated data for SLKD.

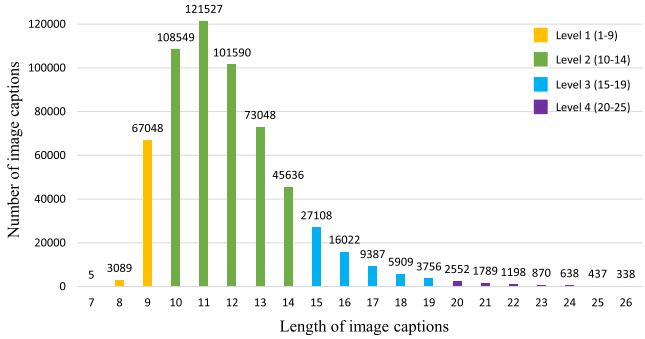


Fig. 5. Length distribution of image captions in MS COCO train set. According to the division of the four length level plan in Section III-B, most (88%) image captions are concentrated in the shorter length level 1 and 2, while the longest level 4 only accounts for 1.3%.

IV. Experiments

A. Dataset and Metrics

To evaluate the effectiveness of our method, we conduct experiments on the popular MS COCO dataset [53], which contains 123,287 images with at least 5 ground-truth captions for each image. We follow the data split setting as in [54], where 113,287, 5,000, and 5,000 images are used for training, validation, and testing, respectively. We further analyze the length distribution of the image captions in the MS COCO training set, as shown in Fig. 5. The data was collected with a minimal caption length of 8, resulting in non-uniform length distribution.

To evaluate the quality of the generated captions, we use standard metrics, including BLEU [55], ROUGE [56], METEOR [57], CIDEr [18], and SPICE [58]. All these metrics except SPICE calculate the similarity between the reference and candidate image captions by considering their n-grams similarity. On the other hand, SPICE is based on scene-graph synonym matching which considers a scene-graph representation of an

TABLE I
PERFORMANCE OF THE LENGTH-CONTROLLABLE AOANET AND VLP ON MS COCO KARPATHY'S `test` SPLIT

Metrics		SPICE	CIDEr	METEOR	BLEU@4	SPICE	CIDEr	METEOR	BLEU@4
Models		AoANet				VLP			
Original Results		21.3	118.4	28.3	**36.9**	21.2	116.9	28.5	**36.5**
4-Level	Lv1	19.6	107.4	25.9	33.1	18.9	103.0	25.2	31.8
	Lv2	21.7	117.6	28.6	35.8	21.4	**118.7**	28.8	36.0
	Lv3	22.7	79.9	28.7	26.6	22.4	92.5	**29.3**	28.4
	Lv4	22.7	29.5	27.7	20.2	22.4	40.0	28.5	21.9
5-Level	Lv1	19.7	108.7	26.0	33.5	18.7	101.0	25.0	30.9
	Lv2	21.6	**118.8**	28.5	36.1	21.2	117.3	28.4	35.9
	Lv3	22.6	92.9	**29.0**	28.7	22.3	100.5	**29.3**	29.9
	Lv4	**23.0**	48.4	28.2	22.7	22.4	60.4	28.7	24.0
	Lv5	22.9	18.9	27.2	18.8	**22.5**	28.1	28.1	20.3

The original results of AoANet and VLP are obtained from models trained by ourselves with the official codes and settings provided by the authors. All values are reported in percentage (%).
The Bold entities indicate the best results.

image by encoding objects, attributes, and relations. According to [58], [59], SPICE and METEOR correlate best with human judgments in terms of caption quality among all these metrics. Moreover, since most ground-truth image captions in the `test` splits are short, the performance of n-gram-based metrics can be negatively affected when evaluating long candidate captions (e.g., CIDEr contains a length penalty term). Fortunately, SPICE is robust to the length of candidate captions, thus it should be the prior metric for the evaluation of long captions.

To evaluate the diversity of the generated captions, we sample the same number of image captions for each model, and use Div-1, Div-2, and Self-CIDEr [60] for evaluation. Div-1/2 computes the ratio of distinct uni/bi-grams in generated captions to the total number of words in the caption set, respectively. Self-CIDEr [60] is a recently proposed metric that focuses on semantic diversity. The higher these values the more diverse the captions are.

B. Implementation Details

For AoANet, VLP, M^2 Transformer, X-LAN, and their length-controllable variants, we adopt their official codes and settings for training, inference, and evaluation. For LaNAR models, we consider two variants: 1) LaNAR-BERT, which is a BERT-like model with 12 layers, 12 attention heads, and a hidden size of 768. We initialize it from a pre-trained BERT-base [12] model. 2) LaNAR-Transformer, which is an encoder-decoder-based transformer model, with a 6-layer encoder and 3-layer decoder. The hidden size and attention heads are set to 512 and 8, respectively. We represent each input image as 100 object proposals extracted by a Faster RCNN [61] pre-trained on the Visual Genome [62] dataset. We take the intermediate results at the `fc6` layer (2048-D) of the Faster RCNN as the region features $\boldsymbol{F}_e$. The classification labels $\boldsymbol{F}_c$ containing 1,600 object categories are obtained from the final `softmax` layer. The localization feature of each proposal is a 5-tuple containing the normalized coordinates of the top-left and bottom-right corners of the proposal and its relative area to the whole image.

We train all LaNAR models for 40 epochs with a batch size of 256, using the AdamW [63] optimizer with a weight decay of 1e-2. The learning rate is linearly warmed up from 0 to 5e-5 during the first 1,000 iterations and is then cosine decayed to 0 for the remained iterations. We use a label smoothing of 0.1, and a gradient clipping threshold of 1.0. When applying the REST scheme on LaNAR models, we initialize the models from the cross-entropy trained models and further finetune them for 25 epochs with a batch size of 64 and a learning rate of 2e-5. When using SLKD for LaNAR models, we obtain the generated data using a pre-trained 4-Level VLP model as the teacher, which generates 5 captions on each length level through beam search, resulting in 20 generated captions for each image. Then, we subsample a generated dataset with the same number of training samples as the original dataset and with the identical length-level distribution for each training sample. We merge the generated data with the original data with a 1:1 ratio.

C. Performance on Auto-Regressive Models

We show the performance of our length-controllable auto-regressive models in Tables I and II. From Table I, in the length range `[10,14]` where the reference captions in MS COCO are mostly distributed, our 4-Level and 5-Level versions of length-controllable VLP [6] and AoANet [4] both achieve competitive or better performance than the original results. Our 4-Level VLP even outperforms the original VLP by 1.8% in terms of the CIDEr score. This indicates that our length-controllable models can maintain or even boost the performance of the original models on a normal length range. On longer length ranges, we find n-gram-based metrics like CIDEr drops severely. However, as we discussed in Section IV-A, this does not mean the captions generated on these levels are poor in quality. From the example in Fig. 1, the 4-Level VLP generates high-quality image captions on all length levels. Specifically, on the shortest level, the image is concisely described from a global perspective, ignoring many important details, while on the longest level, 4-Level VLP covers all the fine-grained visual concepts in the image, such as "pitcher throws a pitch", "batter up to plate " and "catcher ready to catch the ball", some are even missed in the reference captions. This is also supported by more visualization results in Fig. 9. Moreover, our models generally achieve remarkable SPICE scores on longer length levels, i.e., the 5-Level AoANet achieves 23.0 and 22.9 SPICE scores on levels 4 and 5, respectively, which are more than 1.6% higher than the original result.

TABLE II
PERFORMANCE OF THE LENGTH-CONTROLLABLE AR MODELS WITH CIDER OPTIMIZATION ON MSCOCO KARPATHY'S `test` SPLIT

Models	vanilla Transformer			M^2 Transformer			X-LAN		
	SPICE	CIDEr	METEOR	SPICE	CIDEr	METEOR	SPICE	CIDEr	METEOR
Original Results	22.6	129.5	29.1	22.6	131.2	29.2	23.4	132.0	29.5
Length-controllable									
Lv1	20.1	113.4	26.0	20.5	114.9	26.2	20.4	115.6	26.6
Lv2	22.3	**130.6**	29.2	22.7	**133.0**	29.4	22.4	**132.5**	**29.6**
Lv3	23.0	101.1	29.2	23.4	102.5	**29.7**	23.5	102.8	**29.6**
Lv4	23.4	58.5	**29.3**	23.8	59.4	29.5	23.9	59.4	29.5
Lv5	**23.5**	31.2	28.7	**24.0**	32.0	29.0	**23.8**	31.8	28.9

The Bold entities indicate the best results.

TABLE III
PERFORMANCE OF LANAR MODELS ON MS COCO KARPATHY'S `test` SPLIT

Methods	Teacher Forcing				REST + SLKD			
	SPICE	CIDEr	METEOR	BLEU@4	SPICE	CIDEr	METEOR	BLEU@4
MIR [43]	20.6	109.5	27.2	32.5	-	-	-	-
CMAL [44]	-	-	-	-	21.8	124.0	28.1	37.3
Single Level LaNAR-BERT	21.7	116.8	27.9	35.0	22.8	127.8	29.0	38.5
Single Level LaNAR-Transformer	21.7	117.2	28.0	35.3	22.9	128.6	28.9	38.3
LaNAR-BERT Lv1	19.5	101.6	25.4	30.0	20.1	110.4	25.9	34.3
LaNAR-BERT Lv2	21.8	118.4	28.6	34.7	23.1	**130.1**	29.3	**39.7**
LaNAR-BERT Lv3	**22.3**	90.5	28.6	26.8	24.0	101.3	**29.8**	31.2
LaNAR-BERT Lv4	22.2	39.9	27.7	19.9	**24.1**	55.9	29.7	25.3
LaNAR-Transformer Lv1	19.6	102.1	25.6	30.5	20.1	108.9	25.8	33.7
LaNAR-Transformer Lv2	22.1	**119.3**	28.7	**35.6**	22.9	130.0	29.4	39.5
LaNAR-Transformer Lv3	22.1	90.9	**28.8**	27.4	23.8	101.7	29.6	31.1
LaNAR-Transformer Lv4	22.2	41.8	28.0	21.1	24.0	56.6	29.6	25.3

The Bold entities indicate the best results.

In Table II, we investigate three length-controllable AR models for SCST optimization, including the vanilla Transformer [11], M^2 Transformer [9], and X-LAN [10]. From the table, we obtain similar observations as above, where the length-controllable models achieve SOTA performance on the mostly-distributed length level, and significantly improve the SPICE score on long length levels. These results demonstrate the remarkable performance and generalization ability of the proposed length level embedding in existing AR image captioning models, whether using an LSTM-based decoder or Transformer-based decoder, and whether using the Teacher Forcing scheme or SCST scheme.

D. Performance of LaNAR Models

Here, we evaluate the performance of our proposed LaNAR captioning paradigm. The number of refine steps for 4-level models is set to 10, 15, 20, and 25, for levels 1-4, respectively, so that we can compare the LaNAR models with the autoregressive models under roughly the same decoding complexity. From the table, our LaNAR models outperform previous NAR captioning models like MIR [43] and CMAL [44]. We also implement a single-level version of LaNAR models, where the length range is set to `[1,25]` and the number of decoding steps is set to 25. The results are shown in Table III. Compared with single-level LaNAR models, 4-Level LaNAR models achieve clearly better performance on all metrics on level 2, and yield significantly higher SPICE scores on level 3 and level 4, which coincides with the observations in Section IV-C. Moreover, when adopting SLKD and the proposed REST scheme during training, the LaNAR models achieve competitive performance to the AR baselines in Table II. These results demonstrate the effectiveness of our length-controllable approach in the non-autoregressive image caption model.

We further evaluate the speed advantage of the LaNAR models, where we vary the number of refine steps T from 10 to 25 for caption generation on the fourth level of the LaNAR-BERT model. As shown in Fig. 6, our LaNAR-BERT can use a smaller $T = 20$ to achieve comparable results with 4-Level VLP. We can acquire further speedup ($2.1\times$) by setting $T = 12$, with a small sacrifice on SPICE (0.3%). We also evaluate the runtime speed of LaNAR-BERT. On one NVIDIA RTX 3090 GPU, one forward pass of LaNAR-BERT takes 2.9 ms. When using 10 refinement steps for all length levels, LaNAR-BERT requires 32 ms, 36 ms, 41 ms, and 43 ms, for generating one caption on length-level 1 to 4, respectively. On the other hand, the 4-Level VLP model, which has a similar network architecture as LaNAR-BERT, requires 77 ms to generate a caption with 25 tokens, which is $1.8\times$ longer than LaNAR-BERT. Nevertheless, the performance obtained by LaNAR-BERT with 10 refine steps is still competitive with the performance of the 4-Level VLP, which verifies the capability of LaNAR-BERT for efficient image captioning decoding.

E. Performance Analysis of LaNAR Models

1) Ablation Studies on REST: In Table IV , we provide the results of some ablation studies on the proposed REST scheme. From the results, the REST scheme significantly improves the

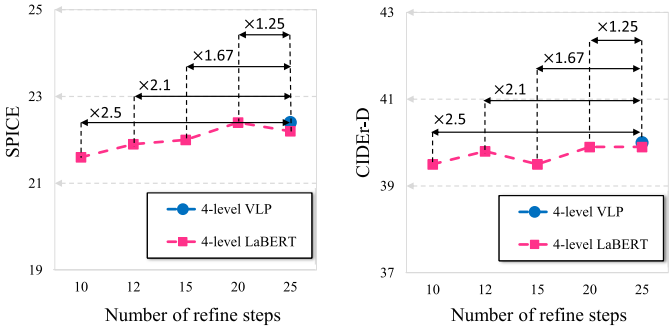


Fig. 6. Performance of LaNAR-BERT on the fourth length level with different numbers of decoding steps.

TABLE IV
PERFORMANCE ANALYSIS FOR 4-LEVEL LANAR-BERT ON MS COCO KARPATHY'S `test` SPLIT

Models	SPICE	CIDEr	METEOR
LaNAR-BERT	21.8	118.4	28.6
LaNAR-BERT w/ SLKD & REST	**23.1**	**130.1**	**29.3**
LaNAR-BERT w/ Original SLKD	20.9	113.9	27.2
+ subsampling	21.5	117.1	27.9
+ merge original data	**22.3**	**121.4**	**28.6**
LaNAR-BERT w/ REST	**22.7**	**128.6**	**29.1**
w/o SC reward	22.5	125.9	28.9
w/o RE reward	22.1	124.6	28.9
$\alpha = 0.0$	21.7	116.6	28.2
$\alpha = 0.5$	21.8	118.2	28.4
$\alpha = 0.8$	21.8	**118.4**	**28.6**
$\alpha = 1.0$	**22.0**	118.2	28.4
$\gamma = 0.80$	**21.8**	**118.4**	**28.6**
$\gamma = 0.90$	21.4	118.3	28.4
$\gamma = 1.0$	21.0	116.0	27.8
$\omega = 0.3$	**22.4**	117.4	28.5
$\omega = 0.5$	21.8	**118.4**	**28.6**
$\omega = 0.7$	22.1	117.5	28.4
$\omega = 1.0$	21.9	117.8	28.2

For simplicity, only results on level-2 (10-14) are shown.
The Bold entities indicate the best results.

performance of LaNAR-BERT, making it comparable to SOTA methods under the CIDEr optimization setting. We also show the importance of the different components in the reward function of REST, i.e., the self-critical (SC) term, and the refinement-enhanced (RE) term. See the results in Table IV. From the table, after removing the SC term or RE term in (12), the CIDEr score of LaNAR-BERT w/ REST drops by 2.7% and 4.0%, respectively, indicating that both two terms play a critical role in our REST scheme.

2) Discussions on SLKD: To analyze the effect of the ratio of the generated data in the merged dataset during SLKD, we gradually increase the proportion of generated data from 0.0 to 1.0 in the merged dataset. We train our LaNAR-BERT model on the new dataset and present the results in Fig. 7. From the figure, the performance of LaNAR-BERT generally increases along with the proportion of the generated data. Saturation may be observed when the proportion is higher than 0.5. Setting the default ratio of the generated data to 0.5, we evaluate the importance of our modifications to the sequence-level knowledge distillation (SLKD) on LaNAR-BERT. The results are presented in Table IV. From the results, we find that the original SLKD degrades the performance of LaNAR-BERT. By subsampling the generated caption data to have the same length-level distribution as the original dataset and merging the original and generated data together for training, we obtain a clear performance boost on LaNAR-BERT, where the CIDEr score on level 2 is improved by 3.0%. Moreover, compared with the results in Table I, LaNAR-BERT with SLKD also outperforms the autoregressive models (4-Level AoANet and 4-Level VLP) on all metrics on level 2. These results have verified the effectiveness of our modifications to the SLKD training scheme.

3) Hyper-Parameter Analysis: We analyze the effect of several key hyper-parameters in LaNAR-BERT, including the `[EOS]` decay factor γ in (10), the balance weight ω in (8), and the global update factor α in (11). By default, we set $\alpha = 0.8$, $\gamma = 0.8$, and $\omega = 0.5$, where LaNAR-BERT achieves the best performance under a normal training setting (w/o SLKD & REST). Then, we change the value of one of these hyper-parameters while keeping the other two fixed. As shown in Table IV, after turning off the global update rule, i.e., $\alpha = 0$, the CIDEr score of LaNAR-BERT on the second level drops by 1.8%. Removing the `[EOS]` decay ($\gamma = 1$) also decrease the performance by 2.4%. Besides, choosing a proper value of the balance weight ω is also important, where we obtain a gain of 0.6% when decreasing it from 1.0 to 0.5.

TABLE V
PERFORMANCE OF THE ORACLE RERANKING AND THE ADAPTIVE LENGTH LEVEL RERANKING FOR LENGTH-CONTROLLABLE AR MODELS AND LANAR-BERT ON MS COCO KARPATHY'S `test` SPLIT

Models	SPICE	CIDEr	METEOR	BLEU@4
4-Level VLP Oracle	28.4	137.8	34.9	38.7
4-Level VLP w/ LLRT	23.2	122.6	29.4	37.5
4-Level AoANet Oracle	28.0	135.7	34.6	38.3
4-Level AoANet w/ LLRT	21.9	119.5	28.9	36.9
LaNAR-BERT Oracle	28.2	137.3	35.1	36.4
LaNAR-BERT w/ LLRT	23.1	122.8	29.1	35.3
Most frequent length level	21.1	113.8	27.6	34.5
Average length level	21.2	112.0	27.6	33.5
CIDEr score regression	21.5	117.7	28.2	34.5
Training from scratch	22.3	121.1	28.8	35.1
CLIPScore reranking	21.5	117.6	28.4	34.2

F. Performance of LLRT

In this section, we evaluate the performance of the proposed length level reranking transformer (LLRT) on 4-Level VLP, 4-Level AoANet, and 4-Level LaNAR-BERT. Specifically, we first obtain their oracle reranking performance, where for each image we compute the evaluation score for the generated caption on each length level and adopt the highest score for calculating the whole-dataset performance. As shown in Table V, all our length-controllable models achieve strong performance under the oracle evaluations, showing the good complementarity of the captions generated on different length levels. Moreover, we show the performance of the length-controllable models reranked by

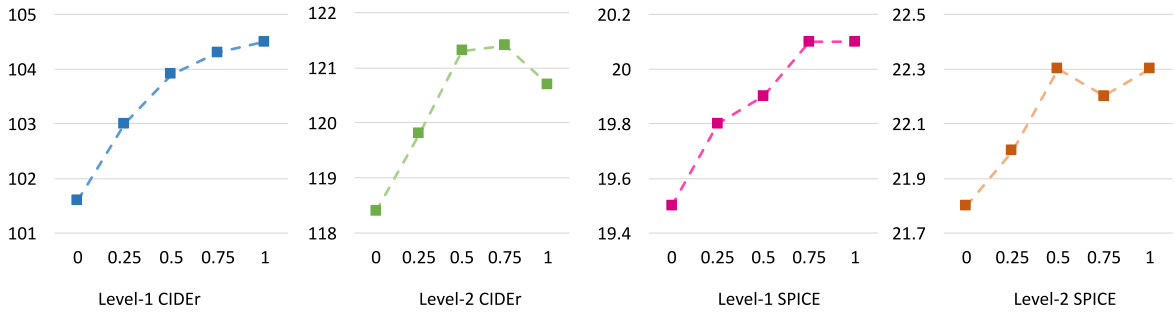


Fig. 7. Performance w.r.t. the proportion of generated data in SLKD.

the proposed LLRT. Compared with the results in Tables I and III, the adaptive reranking performance of 4-Level VLP, 4-Level AoANet, and 4-Level LaNAR-BERT outperforms their results before adaptive reranking on all levels and metrics by a large margin, and also significantly outperforms the results of the original VLP and AoANet. This demonstrates the effectiveness of the proposed LLRT. Nevertheless, there still exists a large gap between the adaptive reranking and the oracle reranking performances.

Further, we show the performance of several other design choices for the LLRT model, including 1) using the most frequent length level in the reference captions of an image as its ground-truth length level, and training a transformer model to directly predict the length level from the image feature through the cross-entropy loss; 2) similar to the first option, but adopt the average length of the reference captions of an image to determine its ground-truth length level; 3) train the LLRT model to directly regress the CIDEr score of the captions; 4) randomly initialize the LLRT model, instead of using the pre-trained weights of length-controllable VLP or LaNAR-BERT; and 5) using a reference-free metric, i.e., CLIPScore [46], for length-level reranking.

From the table, directly predicting the most suitable length level from the image as in the first two options leads to inferior performances. This may be due to two reasons. First, the most frequent or average length level could be a bad indicator of the most suitable length level of an image. Second, the implicit relationship between the semantic complexity of the image and the caption cannot be simply inferred from the image. Besides, CIDEr regression also clearly degrades the performance of LLRT, we hypothesize that it is difficult to precisely estimate the CIDEr score of the generated captions independently. On the other hand, the proposed LLRT estimates the relative quality of the captions for all length levels jointly through a scoring head, leading to better performance. Moreover, LLRT applies a Softmax function on the scores of all length levels, which introduces competition into the learning process, and thus may be beneficial for ranking purposes. We also find that the weight initialization is beneficial to the training of LLRT, where training LLRT from scratch drops the performance slightly. Lastly, using CLIPScore leads to much lower performance than using the proposed LLRT. This result shows that the proposed LLRT is able to mimic human judgments on the semantic complexities of the images thanks to the use of reference-based optimization during training. Reference-free metrics (such as CLIPScore [46], UMIC [48], and VIFIDEL [47]), however, may not be able to provide this guidance and thus may fail to find the proper length level for an image.

TABLE VI
PERFORMANCE ON MSCOCO KARPATHY'S `test` SPLIT

Models	SPICE	CIDEr	METEOR
Vision-language pre-training			
OFA-Large [64]	26.2	150.7	32.2
BLIP2-OPT-2.7B [65]	-	145.8	-
BLIP2-OPT-6.7B [65]	-	145.2	-
LEMON-Huge [66]	25.5	145.5	31.4
CoCa [67]	24.7	143.6	33.9
SimVLM-Huge [68]	25.4	143.3	33.7
VinVL-Large [69]	25.1	140.9	31.1
X-VLM [70]	-	140.8	-
Strong vision backbones			
GRIT [71] (Swin-L)	24.2	**142.2**	30.5
ExpansionNet [72] (Swin-L)	24.5	140.4	30.3
PureT [73] (Swin-L)	24.2	138.2	30.2
RSTNet [74] (ResNeXt-152)	23.3	135.6	29.8
LaNAR-BERT (Swin-L) w/ REST Lv1	20.5	116.9	26.7
LaNAR-BERT (Swin-L) w/ REST Lv2	24.7	140.5	30.4
LaNAR-BERT (Swin-L) w/ REST Lv3	25.3	107.9	**30.7**
LaNAR-BERT (Swin-L) w/ REST Lv4	**25.4**	60.1	30.6
LaNAR-BERT (CLIP-ViT-L) w/ REST Lv1	20.9	117.2	27.2
LaNAR-BERT (CLIP-ViT-L) w/ REST Lv2	24.6	139.6	30.5
LaNAR-BERT (CLIP-ViT-L) w/ REST Lv3	25.1	105.5	30.7
LaNAR-BERT (CLIP-ViT-L) w/ REST Lv4	25.1	58.9	30.4
W/o pre-training & strong backbones			
X-LAN [10]	23.4	132.0	29.5
M^2Transformer [9]	22.6	131.2	29.2
TCIC [75]	22.4	132.9	29.2
GET [76]	22.8	131.6	29.3
DLCT [77]	23.0	**133.8**	29.5
AoANet [4]	22.4	129.8	29.2
VLP [6]	23.2	129.3	29.3
LaNAR-BERT w/ ALL	**23.5**	133.6	**29.7**
LaNAR-Transformer w/ ALL	23.4	132.8	29.5

LaNAR model w/ All denotes using SLKD, REST, and LLRT to boost the performance of LaNAR models.
The Bold entities indicate the best results.

G. Comparisons With State-of-the-Arts

Recent SOTA image captioning models have achieved remarkable performance on the MSCOCO dataset with the help of large-scale vision-language pre-training or strong vision

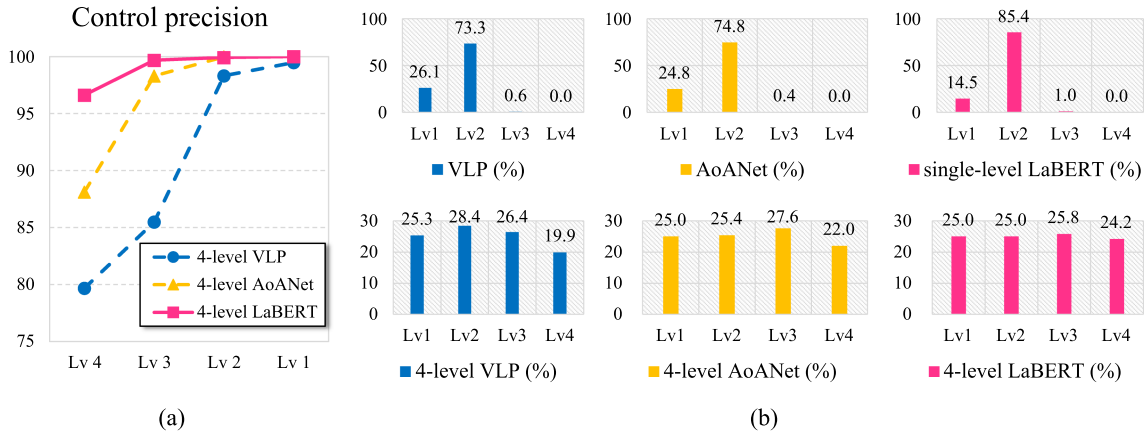


Fig. 8. Analysis of controllability and diversity on `test` split. (a) Control precision of our 4-level version of AoANet, VLP, and LaNAR-BERT. (b) Length distributions of image captions generated by our 4-level length-aware models and their counterparts.

TABLE VII
PERFORMANCE OF LANAR-BERT ON CONCEPTUAL CAPTIONS VALIDATION SET

Models	SPICE	CIDEr	METEOR	Average Length
VLP	15.5	67.0	9.7	-
LaNAR-BERT				
Lv1	15.2	66.9	9.4	8.4
Lv2	15.7	67.3	9.6	10.5
Lv3	15.8	65.2	9.7	13.5
Lv4	16.1	53.5	9.6	17.8
Lv5	16.0	20.9	9.3	22.1
Lv6	15.9	10.8	9.1	27.5

TABLE VIII
DIVERSITY ANALYSIS. BS DENOTES BEAM SEARCH

Models	AoANet		VLP		LaNAR-BERT
	BS	4-Level	BS	4-Level	4-Level
SelfCIDEr [60]	0.590	0.689	0.623	0.762	**0.841**
Div-1	0.291	0.378	0.313	0.406	**0.411**
Div-2	0.462	0.523	0.470	0.559	**0.575**

The Bold entities indicate the best results.

backbones like Swin Transformer [78] pre-trained on ImageNet22k [79]. To better demonstrate the effectiveness of our method, in Table VI, we provide a more detailed comparison under different settings.

We first show the performance of LaNAR-BERT with strong vision backbones, i.e., Swin-Transformer Large (Swin-L), and CLIP Vision Transformer Large (CLIP-ViT-Large). From the table, LaNAR-BERT achieves competitive performance to the SOTA baselines with the same backbones, while enjoying length-controllable and non-autoregressive decoding. Moreover, in the standard setting, i.e., without using large-scale vision-language pre-training or strong vision backbones, the LaNAR-BERT w/ REST & SLKD & LLRT model achieves superior performance than existing state-of-the-art models, which demonstrates the effectiveness of our proposed methods.

We also perform experiments on Conceptual Captions [80], a large-scale and more challenging dataset with extremely diverse semantics and a large variance in caption length. We divide the captions in Conceptual Captions into 6 levels according to the length distribution: `[1,8]`, `[9,12]`, `[13,16]`, `[17,20]`, `[21,28]`, and `[29,44]`. Then, we train a length-controllable LaNAR-BERT with this new division strategy following the training and evaluation settings in VLP [6] on Conceptual Captions, the results are shown in Table VII. From the table, LaNAR-BERT trained on Conceptual Captions successfully controls the length of the captions on all 6 levels. Moreover, on short levels (Level-1 to Level-3), LaNAR-BERT achieves competitive performance with VLP in terms of CIDEr, SPICE, and METEOR; while on longer length levels, LaNAR-BERT is superior in terms of SPICE. These results are aligned with our observations on the MSCOCO dataset, showing that our LaNAR paradigm can be applied to more challenging datasets.

H. Controllability and Diversity Analysis

In this section, we further analyze the "control precision" of the length level embedding, i.e., given a length level embedding, the probability of generating image captions within the desired length range. We calculate the control precision for the 4-level version of AoANet, VLP, and LaNAR-BERT, and present the results in Fig. 8(a). As shown in the figure, all methods accurately control the length of the generated image captions, and our non-autoregressive model, LaNAR-BERT, yields the best control precision (more than 95%) among all levels. This result verifies the effectiveness of the proposed length-level embedding in generating length-controllable image captions. Besides, the control precision drops on longer levels, which may be due to the lack of long captions in the MS COCO dataset.

We also perform diversity analysis for the image captions generated by different models, as shown in Fig. 8(b) and Table VIII. From Fig. 8(b), the length of the image captions generated by our length-aware models are uniformly distributed among all length levels. On the contrary, the results of the original AoANet, VLP, and the single-level LaNAR-BERT distribute mainly in the shortest two levels. We further evaluate the diversity of the image captions on n-gram diversity metrics like Div-1 and Div-2, as well as the recently proposed SelfCIDEr [60] score

Ori	A red airplane sitting on top of an airport runway.	VLP
Lv1	A person standing next to a small plane.	
Lv2	A red and white plane sitting on top of a runway.	
Lv3	A red and white plane sitting on top of an airport tarmac.	
Lv4	A red and white plane sitting on top of a tarmac with a man standing next to it .	
Ori	A red and white plane sitting on a runway.	AoANet
Lv1	A small red plane on a run way.	
Lv2	A red and white plane sitting on a runway.	
Lv3	A red and white plane sitting on top of a runway next to a runway.	
Lv4	A red and white airplane with a red tail on the tail and a person standing in front of it.	
Lv1	A small red airplane sitting on a runway.	LaBERT
Lv2	A small red and white airplane sitting on a runway.	
Lv3	A red and white plane parked on a runway with a person beside it.	
Lv4	A red and white plane sitting on top of a tarmac with a person standing next to it.	
Lv1	A red airplane sitting on a runway.	LaBERT+SLKD
Lv2	A red and white plane sitting on the runway.	
Lv3	A red and white plane sitting on the with a person standing on it.	
Lv4	A red and white plane sitting at an airport with a person standing next to it.	
Lv1	A red airplane parked on a runway.	LaBERT+REST
Lv2	A red plane parked on a runway on a grassy field.	
Lv3	A red plane parked on a green field with a person standing by.	
Lv4	A red and white plane is sitting on a grassy field at an airport next to a runway.	

Ori	A man riding on the back of a brown horse.	VLP
Lv1	A man riding a horse down a street.	
Lv2	A man riding on the back of a brown horse.	
Lv3	A man riding on the back of a brown horse down a city street.	
Lv4	A man in a green jacket riding a brown horse in front of a group of people on a sidewalk.	
Ori	Two men and a woman are riding horses.	AoANet
Lv1	A group of people sitting on some horses.	
Lv2	A group of people riding on the backs of horses.	
Lv3	A group of people riding on the back of horses next to a building	
Lv4	A group of people sitting on horses in front of a building with a group of people standing around.	
Lv1	A man that is sitting on a horse.	LaBERT
Lv2	A man riding on the back of a brown horse.	
Lv3	A man riding on the back of a brown horse next to other people.	
Lv4	A man in a green coat and hat sitting on a horse with a woman standing next to him.	
Lv1	A group of people sitting on horses.	LaBERT+SLKD
Lv2	A group of people riding on the back of horses.	
Lv3	A man riding on the back of a horse in a crowd of people.	
Lv4	A man riding on the back of a horse next to a little girl in front of a building.	
Lv1	A group of people riding horses on a street.	LaBERT+REST
Lv2	A group of people riding horses down a crowded street.	
Lv3	Two man riding brown horses with a lot of people standing next to it.	
Lv4	Two men in green suit riding on the back of brown horses next to a woman on a city street.	

Ori	A group of zebra standing next to each other.	VLP
Lv1	Three zebras are standing near water.	
Lv2	A group of zebra standing next to each other.	
Lv3	A herd of zebra standing next to each other on a field.	
Lv4	A herd of zebra standing on top of a lush green field next to a watering hole.	
Ori	A group of zebras standing next to a watering hole.	AoANet
Lv1	A group of zebras standing near a water.	
Lv2	A herd of zebra standing next to each other on a field.	
Lv3	A herd of zebra standing next to each other on top of a dirt field.	
Lv4	A herd of zebra standing next to each other on a dirt field next to a body of water.	
Lv1	A group of zebra standing near a lake.	LaBERT
Lv2	A group of zebras standing next to a watering hole.	
Lv3	A group of zebras standing next to a herd of antelope.	
Lv4	Three zebras standing next to a body of water with a couple of other animals in the background.	
Lv1	A group of zebras standing in the water.	LaBERT+SLKD
Lv2	A group of zebras standing next to a watering hole.	
Lv3	A group of zebras and a gazelle standing by a watering hole.	
Lv4	Three zebras and a gazelle are standing in the water at the edgeof a watering hole.	
Lv1	A group of zebras standing by a water.	LaBERT+REST
Lv2	A group of zebras standing next to a watering hole.	
Lv3	A herd of zebras and antelopes are standing next to each other.	
Lv4	A herd of zebras and antelopes are standing next to each other next to a watering hole.	

Fig. 9. Examples of length-controllable image captioning from Karpathy's `test` split. "Ori" denotes the original results of AoANet or VLP. "Lv n" denotes results on the n th length level.

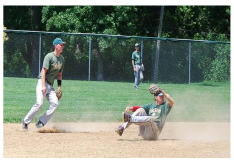
A man on a surfboard is riding a wave.

A bunch of oranges on a plate.

A dog carrying a yellow frisbee.

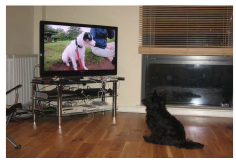
Two tennis rackets sitting on a chair.

A close up of a computer keyboard.



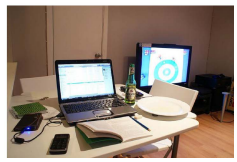
A baseball player is sliding to a base while another player is trying to catch it in the air.

A white plate with a cut in half sandwich on it and a spoon sitting on top of it.



A dog is sitting on the floor watching a dog on a flat screen television.

A hand holding a cell phone with a tv screen showing a picture of two cows in the background.



A laptop computer sitting on a desk with a tablet with a cell phone and pen next to it.

Fig. 10. Sampled results of length level reranking on LaNAR-BERT from MS COCO Karpathy's `test` split. Similar to the example in Fig. 1(b), short and precise captions are preferred for semantically simple images (first row), while long and informative captions are selected when describing more complex images (second row).

that focuses on semantic diversity. From Table VIII, our 4-level models perform clearly better on all metrics, which means we can obtain diverse captions for an image with our length-aware image captioning models. Interestingly, our non-autoregressive model LaNAR-BERT significantly outperforms all compared autoregressive methods on all three diversity metrics.

V. Qualitative Results

In this section, we show some examples of the image captions generated by our length-controllable models. As shown in Fig. 9, in general, our length-controllable models are able to correctly describe the image, while also controlling the length of the generated captions within the desired length range. More specifically, the long captions (level 3 and level 4) tend to contain more visual concepts, while the short captions (level 1 and level 2) describe the image briefly.

Moreover, In Fig. 10, we present some examples of the length level reranking results on 4-Level LaNAR-BERT. From the figure, for pictures with simple backgrounds and a few foreground objects, the model chooses to produce short captions, as shown in the first row. On the contrary, for pictures with complex scenes, the model generates long sentences to describe the visual information in detail, as shown in the second row. This demonstrates that our model has the ability to find out a suitable caption length when depicting a given image.

VI. Conclusion

In this paper, we propose to use a length-level embedding for length-controllable image captioning. By simply adding our length level embedding on the word embeddings of input tokens, we endow existing image captioning methods with the ability to control the length of their predictions. Besides, to automatically determine the most suitable length level for an image, we propose to learn a length level reranking transformer through reinforcement learning, so as to capture the implicit relationship between the semantic complexity of the image and the language description. Furthermore, to improve the decoding efficiency of long captions, we propose a non-autoregressive image captioning paradigm, LaNAR, that generates image captions in a length-irrelevant complexity. We further develop a sequence-level knowledge distillation strategy as well as a refinement-enhanced sequence training scheme for LaNAR to boost its performance. In the experiments, our length-aware models generate high-quality and length-controllable image captions, and our length-level reranking transformer consistently improves the final performance. Moreover, our LaNAR models not only achieve comparable performance with the SOTA autoregressive methods in a much smaller computational complexity, but also perform better than the autoregressive baselines in terms of controllability and output diversity.

References

[1] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.

[2] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3156–3164.

[3] Q. Wu, C. Shen, P. Wang, A. Dick, and A. van den Hengel, "Image captioning and visual question answering based on attributes and external knowledge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1367–1381, Jun. 2018.

[4] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 4633–4642.

[5] G. Li, L. Zhu, P. Liu, and Y. Yang, "Entangled transformer for image captioning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 8928–8937.

[6] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and VQA," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, pp. 13041–13049, 2020.

[7] X. Chen et al., "Microsoft COCO captions: Data collection and evaluation server," 2015, *arXiv:1504.00325*.

[8] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7008–7024.

[9] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10578–10587.

[10] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10971–10980.

[11] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 4171–4186.

[13] Y. Kim and A. M. Rush, "Sequence-level knowledge distillation," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2016, pp. 1317–1327.

[14] C. Deng, N. Ding, M. Tan, and Q. Wu, "Length-controllable image captioning," in *Proc. 16th Eur. Conf.*, Glasgow, UK, Aug. 23–28, 2020, pp. 712–729.

[15] K. Xu et al., "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[16] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Representations*, 2015.

[17] P. Anderson et al., "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6077–6086.

[18] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4566–4575.

[19] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4894–4902.

[20] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10685–10694.

[21] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 684–699.

[22] A. Deshpande, J. Aneja, L. Wang, A. G. Schwing, and D. Forsyth, "Fast, diverse and accurate image captioning guided by part-of-speech," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10695–10704.

[23] S. Chen, Q. Jin, P. Wang, and Q. Wu, "Say as you wish: Fine-grained control of image caption generation with abstract scene graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9962–9971.

[24] M. Cornia, L. Baraldi, and R. Cucchiara, "Show, control and tell: A framework for generating controllable and grounded captions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8307–8316.

[25] J. Johnson, A. Karpathy, and L. Fei-Fei, "DenseCap: Fully convolutional localization networks for dense captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4565–4574.

[26] Y. Zheng, Y. Li, and S. Wang, "Intention oriented image captions with guiding objects," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8395–8404.

[27] B. Dai, S. Fidler, R. Urtasun, and D. Lin, "Towards diverse and natural image descriptions via a conditional GAN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2970–2979.

[28] R. Shetty, M. Rohrbach, L. Anne Hendricks, M. Fritz, and B. Schiele, "Speaking the same language: Matching machine to human captions by adversarial training," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4135–4144.

[29] L. Wang, A. Schwing, and S. Lazebnik, "Diverse and accurate image description using a variational auto-encoder with an additive Gaussian encoding space," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5756–5766.

[30] T. Chen et al., ""Factual"or"Emotional": Stylized image captioning with adaptive learning and attention," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 519–535.

[31] A. P. Mathews, L. Xie, and X. He, "SentiCap: Generating image descriptions with sentiments," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 3574–3580.

[32] K. Shuster, S. Humeau, H. Hu, A. Bordes, and J. Weston, "Engaging image captioning via personality," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12516–12526.

[33] C. Gan, Z. Gan, X. He, J. Gao, and L. Deng, "StyleNet: Generating attractive visual captions with styles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3137–3146.

[34] A. Mathews, L. Xie, and X. He, "SemStyle: Learning to generate stylised image captions using unaligned text," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8591–8600.

[35] Y. Kikuchi, G. Neubig, R. Sasano, H. Takamura, and M. Okumura, "Controlling output length in neural encoder-decoders," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2016, pp. 1328–1338.

[36] Y. Liu, Z. Luo, and K. Zhu, "Controlling length in abstractive summarization using a convolutional neural network," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2018, pp. 4110–4119.

[37] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1243–1252.

[38] J. Gu, J. Bradbury, C. Xiong, V. O. Li, and R. Socher, "Non-autoregressive neural machine translation," in *Proc. Int. Conf. Learn. Representations*, 2018.

[39] M. Ghazvininejad, O. Levy, Y. Liu, and L. Zettlemoyer, "Mask-predict: Parallel decoding of conditional masked language models," in *Proc. Conf. Empir. Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 6114–6123.
[40] J. Gu, C. Wang, and J. Zhao, "Levenshtein transformer," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 11179–11189.
[41] M. Stern, W. Chan, J. Kiros, and J. Uszkoreit, "Insertion transformer: Flexible sequence generation via insertion operations," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5976–5985.
[42] C. Wang, J. Zhang, and H. Chen, "Semi-autoregressive neural machine translation," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2018, pp. 479–488.
[43] J. Lee, E. Mansimov, and K. Cho, "Deterministic non-autoregressive neural sequence modeling by iterative refinement," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2018, pp. 1173–1182.
[44] L. Guo, J. Liu, X. Zhu, X. He, J. Jiang, and H. Lu, "Non-autoregressive image captioning with counterfactuals-critical multi-agent learning," in *Proc. 29th Int. Conf. Int. Joint Conf. Artif. Intell.*, 2021, pp. 767–773.
[45] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1171–1179.
[46] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "CLIPScore: A reference-free evaluation metric for image captioning," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 7514–7528.
[47] P. S. Madhyastha, J. Wang, and L. Specia, "VIFIDEL: Evaluating the visual fidelity of image descriptions," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 6539–6550.
[48] H. Lee, S. Yoon, F. Dernoncourt, T. Bui, and K. Jung, "UMIC: An unreferenced metric for image captioning via contrastive learning," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process. (Volume 2: Short Papers)*, 2021, pp. 220–226.
[49] S. Lloyd, "Measures of complexity: A nonexhaustive list," *IEEE Control Syst. Mag.*, vol. 21, no. 4, pp. 7–8, Aug. 2001.
[50] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2000, pp. 1057–1063.
[51] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *Stat*, vol. 1050, pp. 21, 2016.
[52] C. Zhou, J. Gu, and G. Neubig, "Understanding knowledge distillation in non-autoregressive machine translation," in *Proc. Int. Conf. Learn. Representations*, 2019.
[53] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2014, pp. 740–755.
[54] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3128–3137.
[55] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.
[56] C. Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Workshop Text Summarization Branches Out*, 2004, pp. 74–81.
[57] S. Banerjee and A. Lavie, "Meteor: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization*, 2005, pp. 65–72.
[58] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic propositional image caption evaluation," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2016, pp. 382–398.
[59] M. Kilickaya, A. Erdem, N. Ikizler-Cinbis, and E. Erdem, "Re-evaluating automatic metrics for image captioning," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2017, pp. 199–209.
[60] Q. Wang and A. B. Chan, "Describing like humans: On diversity in image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4195–4203.
[61] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[62] R. Krishna et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, 2017.
[63] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in Adam," 2017, *arXiv:1711.05101*.
[64] P. Wang et al., "OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2022, pp. 23318–23340.
[65] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. Int. Conf. Mach. Learn.*, vol. 202, 2023, pp. 19730–19742.
[66] X. Hu et al., "Scaling up vision-language pre-training for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 17980–17989.
[67] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "COCA: Contrastive captioners are image-text foundation models," *Trans. Mach. Learn. Res.*, vol. 2022, 2022.
[68] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, "SimVLM: Simple visual language model pretraining with weak supervision," in *Proc. Int. Conf. Learn. Representations*, 2021.
[69] P. Zhang et al., "VinVL: Revisiting visual representations in vision-language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 5579–5588.
[70] Y. Zeng, X. Zhang, and H. Li, "Multi-grained vision language pre-training: Aligning texts with visual concepts," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2022, pp. 25994–26009.
[71] V.-Q. Nguyen, M. Suganuma, and T. Okatani, "GRIT: Faster and better image captioning transformer using dual visual features," in *Proc. 17th Eur. Conf. Comput. Vis.*, Tel Aviv, Israel, Springer, 2022, pp. 167–184.
[72] J. C. Hu, R. Cavicchioli, and A. Capotondi, "ExpansionNet v2: Block static expansion in fast end to end training for image captioning," 2022, *arXiv:2208.06551*.
[73] Y. Wang, J. Xu, and Y. Sun, "End-to-end transformer based model for image captioning," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 2585–2594.
[74] X. Zhang et al., "RSTNet: Captioning with adaptive attention on visual and non-visual words," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15465–15474.
[75] Z. Fan et al., "TCIC: Theme concepts learning cross language and vision for image captioning," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, 2021, pp. 657–663.
[76] J. Ji et al., "Improving image captioning by leveraging intra-and inter-layer global representation in transformer network," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1655–1663.
[77] Y. Luo et al., "Dual-level collaborative transformer for image captioning," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 2286–2293.
[78] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.
[79] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[80] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2556–2565.

Ning Ding received the ME degree from the School of Software Engineering, South China University of Technology, China, in 2021. She is an algorithm engineer with JD.com, Beijing, China. Her research interests include deep learning in information retrieval and vision-language modeling.

Chaorui Deng is currently working toward the PhD degree with the University of Adelaide, Australia. He is focusing on image/video understanding and vision-language modeling.

Mingkui Tan received the bachelor's degree in environmental science and engineering, the master's degree in control science and engineering, both from Hunan University in Changsha, China, in 2006 and 2009, respectively, and the PhD degree in computer science from Nanyang Technological University, Singapore, in 2014. He is currently a professor with the School of Software Engineering, South China University of Technology, China. From 2014–2016, he worked as a senior research associate on computer vision with the School of Computer Science, University of Adelaide, Australia. His research interests include machine learning, sparse analysis, deep learning, and large-scale optimization.

Qing Du received the BS degree in computer science and technology, the master's degree in computer application, and the PhD degree in computer application from the South China University of Technology, China, in 2002, 2005, and 2014, respectively, where she is currently an associate professor with the School of Software Engineering. Her research interests include information retrieval, recommendation systems, natural language processing, and deep learning.

Zhiwei Ge received the BE and PhD degrees in electronic and information engineering from Tianjin University, in 2007 and 2012, repectively. He is an algorithm engineer with JD.com, Beijing, China. He is focusing on the research, development, and innovation of computer vision and multi-modality search technologies.

Qi Wu received the MSc and PhD degrees in computer science from the University of Bath, U.K., in 2011 and 2015, respectively. He is a senior lecturer (assistant professor) with the University of Adelaide, Australia, and is an associate investigator with the Australia Centre for Robotic Vision (ACRV). He was the ARC Discovery Early Career Researcher Award (DECRA) Fellow between 2019–2021. His educational background is primarily in computer science and mathematics. He works on the Vision and Language problems, including Image Captioning, Visual Question Answering, Visual Dialog etc. His work has been published in prestigious journals and conferences such as *IEEE Transactions on Pattern Analysis and Machine Intelligence*, CVPR, ICCV, AAAI, and ECCV.