Rethinking the Reference-based Distinctive Image Captioning

Yangjun Mao*
Zhejiang University
maoyj0119@zju.edu.cn

Long Chen*
Columbia University
zjuchenlong@gmail.com

Zhihong Jiang
Zhejiang University
zju_jiangzhihong@zju.edu.cn

Dong Zhang
Hong Kong University of Science and Technology, dongz@ust.hk

Zhimeng Zhang
Jian Shao
Zhejiang University

Jun Xiao†
Zhejiang University
junx@zju.edu.cn

ABSTRACT

Distinctive Image Captioning (DIC) — generating distinctive captions that describe the unique details of a target image — has received considerable attention over the last few years. A recent DIC work proposes to generate distinctive captions by comparing the target image with a set of semantic-similar reference images, i.e., reference-based DIC (Ref-DIC). It aims to make the generated captions tell apart the target and reference images. Unfortunately, reference images used by existing Ref-DIC works are easy to distinguish: these reference images only resemble the target image at scene-level and have few common objects, such that a Ref-DIC model can trivially generate distinctive captions even without considering the reference images. For example, if the target image contains objects “towel” and “toilet” while all reference images are without them, then a simple caption “A bathroom with a towel and a toilet” is distinctive enough to tell apart target and reference images. To ensure Ref-DIC models really perceive the unique objects (or attributes) in target images, we first propose two new Ref-DIC benchmarks. Specifically, we design a two-stage matching mechanism, which strictly controls the similarity between the target and reference images at object-/attribute-level (vs. scene-level). Secondly, to generate distinctive captions, we develop a strong Transformer-based Ref-DIC baseline, dubbed as TransDIC. It not only extracts visual features from the target image, but also encodes the differences between objects in the target and reference images. Finally, for more trustworthy benchmarking, we propose a new evaluation metric named DisCIDEr for Ref-DIC, which evaluates both the accuracy and distinctiveness of the generated captions. Experimental results demonstrate that our TransDIC can generate distinctive captions. Besides, it outperforms several state-of-the-art models on the two new benchmarks over different metrics.

*Yangjun Mao and Long Chen are co-first authors with equal contributions to this work. Our code is available on https://github.com/maoyj1998/TransDIC.
†Jun Xiao is the corresponding author.

CCS CONCEPTS
• Computing methodologies → Computer vision.

KEYWORDS
Image Captioning, Distinctiveness, Benchmark, Transformer

1 INTRODUCTION

Image captioning, i.e., generating natural language descriptions to summarize the salient contents of a target image, has drawn much attention from the multimedia community. It has great impacts on many downstream applications, such as helping blind people and developing navigation systems. However, as revealed in [11, 12], conventional image captioning models tend to generate over-generic captions or even identical captions when input images are similar. Obviously, these generic captions neglect the unique details of the target image. Recent captioning works [12, 25, 26, 42] begin to make generated captions more distinctive and ask these captions describe more unique details of each target image, called Distinctive Image Captioning (DIC).

Currently, mainstream DIC works follow the same setting as plain image captioning: using one single image as input, and generating distinctive captions for each image, dubbed as Single-image DIC (Single-DIC). In this setting, they tend to generate a totally distinctive caption. By “totally”, we mean the generated caption is asked to distinguish its corresponding image from all images in the dataset, i.e., dataset-level distinctiveness. To this end, they always resort to reinforcement learning and correspond different rewards [25, 26]. However, this Single-DIC setting has two inherent issues: 1) It is difficult (or impossible) to generate a totally distinctive caption. 2) Even for our humans, we still need some reference images when generating distinctive captions. For example in Figure 1(b), without any reference images, people won’t know what should be emphasized in the target image, and may simply predict “A bathroom with a towel” for the image. In contrast, they will focus on the unique colors of “towel” and “shower curtain”, and predict “A bathroom with a pink towel and a blue shower curtain” when they use the “white shower curtain” in ref-img1 and “yellow towel” in ref-img2 as references.
For human-like distinctive captioning, a recent work proposes to study the DIC task based on a group of semantic-similar reference images, dubbed Reference-based DIC (Ref-DIC). Different from Single-DIC, they use the target image and all reference images as input and these reference images will inform DIC models which parts of the target image should be emphasized. Compared to Single-DIC, the generated captions are only asked to distinguish the target image from the group of reference images, i.e., group-level distinctiveness. Unfortunately, the reference images used in existing Ref-DIC works [43] can be trivially distinguished: these reference images only resemble the target image at the scene-level and have few common objects, thus Ref-DIC models can simply generate distinctive captions even without considering the reference images. For example in Figure 1 (a), target and reference images have no object in common (e.g., "towel", "shower curtain", or "toilet"), each object in target image is unique, such that the Ref-DIC model can trivially generate "a bathroom with a towel" to tell the target and reference images apart.

As mentioned above, reference images are crucial when defining the unique details in the target image. In this paper, we propose two new benchmarks for the Ref-DIC task: COCO-DIC and Flickr30K-DIC. To strictly control the unique details between target and reference images, we propose a two-stage matching mechanism, which can measure image similarity at the object-/attribute-level (vs. scene-level in [43]), and deliberately make target and reference images have some common objects. Under this mechanism, Ref-DIC models can learn to focus on the unique attributes and objects in target image. As the example in Figure 1 (b), compared to "ref-img1", target image has the unique attribute "blue" of "shower curtain" and the unique object "towel".

To achieve group-level distinctiveness, we propose to emphasize both unique attributes and objects in target image. Thus, we also propose a new Transformer-based captioning model named TransDIC, which directly gives each region in target image (target regions) some region references when generating captions. Specifically, we firstly find similar regions from reference images (reference regions) for each target region. Then, we send target region and its corresponding reference regions into the Two-Flow Encoder Module, which consists of a Target flow and a Target-Reference flow. The Target flow aims to encode target image features through self-attention blocks [38]. And the Target-Reference flow enables cross-image interactions between target and reference images through a multi-layer co-attention. Different from the existing Ref-DIC work [43] which proposes an attention module to focus on unique objects in target image, our TransDIC directly enables the feature interactions between target and reference images.

Furthermore, to fully take advantage of ground-truth captions of reference images, we propose a new CIDEr-based [40] metric termed as DisCIDEr. According to our definition of group-level distinctiveness, we believe frequently-used n-grams in ground-truth captions of reference images should be given less weight at evaluation time. The metric can not only directly evaluate the distinctiveness but also preserve the accuracy advantage of CIDEr. Extensive experimental results on multiple Ref-DIC benchmarks (COCO-DIC and Flickr30K-DIC) have demonstrated the effectiveness of our proposed TransDIC.

In summary, we make three contributions in this paper: (1) We proposed two new benchmarks for Ref-DIC, constructed groups in those benchmarks can inform the model which parts in target image should be emphasized. (2) We developed a new model TransDIC for Ref-DIC, which achieves better performance than the state-of-the-art Ref-DIC models in terms of both accuracy and distinctiveness. (3) We proposed a new metric named DisCIDEr which considers both the distinctiveness and accuracy of the generated caption.

2 RELATED WORK

Image Captioning. Most modern image captioning models typically employ an encoder-decoder framework for caption generation [15, 16, 22, 41, 44]. Within this framework, many efforts have been made to improve the architecture, including attention mechanisms [1, 5, 7, 14, 46], graph convolution networks [8, 50, 51], and transformer-based models [10, 17]. Meanwhile, another series of works explore different training objectives at the training stage. For example, Dai et al. [11] and Shetty et al. [35] leverage Generative Adversarial Network (GAN) to improve the diversity of generated captions. Some recent captioning works apply Reinforcement Learning (RL) to captioning and achieve great success [24, 32, 34, 48]. These models directly optimize non-differentiable evaluation metrics (e.g., BLEU [27], CIDEr [40]), which boost the caption generation procedure at the sentence-level.

Distinctive Image Captioning (DIC). Compared with conventional image captioning, DIC is a more challenging task, which tends to generate more informative and descriptive captions. According to the stage they take effect, existing solutions can be coarsely divided into two categories: Inference-based and Training-based methods. Inference-based models mainly modify the caption decoding procedure at inference time and thus can be applied to any
3 PROPOSED BENCHMARKS

In this section, we firstly formally define the Ref-DIC task. Then we describe our solution for Ref-DIC benchmarks construction. Finally, we provide details of our proposed COCO-DIC and Flickr30K-DIC benchmarks for Ref-DIC.

3.1 Task Definition: Reference-based DIC

Given a target image \( I_t \) and a group of \( K \) reference images \( I_r = \{I_r\}_{i=1}^K \), which are semantic-similar to \( I_t \), Ref-DIC models aim to generate a natural language sentence \( S = \{w_1, w_2, \ldots, w_T\} \). The generated sentence \( S \) should not only correctly describe the target image \( I_t \) but also contain sufficient details about \( I_t \), so it can tell apart target and reference images. For example in Figure 1 (b), given the target and reference images, Ref-DIC models aim to generate a distinctive caption “a bathroom with a pink towel, a blue shower curtain and a toilet”. The detail “pink towel” is helpful to distinguish target image from ref-img2 because the “towel” in ref-img2 is “white”. On the contrary, predicting “a bathroom with a towel and a toilet” fails to meet the requirements because it is suitable for both target and reference images.

3.2 Ref-DIC Benchmarks Construction

Given a conventional image captioning dataset \( D \), suppose it contains \( N \) images and each one has \( M \) corresponding ground-truth captions. We build new Ref-DIC benchmarks based on \( D \), by coupling each image (target image) with several semantic-similar reference images. Specifically, each image in \( D \) will be regarded as a target image \( I_t \), and all remaining \( N - 1 \) images are termed as its candidate reference images. For each target image \( I_t \), our goal is to retrieve \( K \) reference images from its candidate reference images to construct the reference image group \( I_r \).

To achieve group-level distinctiveness, \( I_t \) and retrieved \( I_r \) should have some common objects, such that \( I_r \) will inform the model to focus on the unique details in \( I_t \). To this end, we design a two-stage matching mechanism. In the first stage, we construct a coarse-grained group \( I_c \) based on the image-text similarity score for each target image. Then in the second stage, we investigate fine-grained details of \( I_t \) and \( I_c \), and construct a fine-grained group \( I_f \) based on \( I_c \). Finally, we select \( K \) images out of \( I_f \) to construct the \( I_r \). We detailed introduce our two-stage matching pipeline below.

3.2.1 Coarse-grained Group Construction. Following [43], we use an image-text retrieval model to calculate similarity scores between
4.1 Preliminaries

4.1.1 Transformer-based Image Captioning. Transformer [38] follows the standard Encoder-Decoder architecture. It employs the self-attention mechanism to explore the internal correlation within the sequential data, which has been widely adopted by numerous image captioning models [10, 17]. For a given image \(I\), they use proposal features \(I_i\) extracted by an object detector as input: \(X = \{x_i\}_{i=1}^N\) where \(x_i \in \mathbb{R}^d\) is the feature vector for \(i\)-th proposal in \(I\) and \(N\) is the number of proposals. They employ multiple self-attention layers as Encoder, and the outputs of the \(l\)-th layer is calculated as follows:

\[
H_{l-1} = \text{LN} (O_{l-1} + \text{MH}(O_{l-1}, O_{l-1}, O_{l-1})) \quad (1),
\]

\[
O_l = \text{LN}(H_{l-1} + \text{FFN}(H_{l-1})) \quad (2),
\]

where \(O_0\) refers to input proposal features \(X\) and \(O_{l-1}\) is outputs of the \((l-1)\)-th layer. \(\text{LN}(\cdot)\) denotes the layer normalization [2], \(\text{FFN}(\cdot)\) denotes the feed forward network, and \(\text{MH}(\cdot)\) denotes the multi-head attention [38]. Encoded visual features are fed into the Decoder for caption generation. It generates a word probability distribution \(P_t = P(w_t | w_{1:t-1}, I)\) at each step \(t\) conditioning on the previously generated words \(\{w_1, \ldots, w_{t-1}\}\) and image \(I\).

4.1.2 Model Optimization. Mainstream captioning works typically resort to a two-stage procedure for model optimization [10, 34]. Given an image \(I\) and its ground-truth captions \(C = \{c_i\}_{i=1}^M\), they firstly apply a cross-entropy loss (XE) to pre-train the model and then employ reinforcement learning (RL) to finetune sequence generation.

1 Note that the "black helmet" and "black motorcycle" contain no attribute overlap because the attribute "black" belongs to different objects.
When training with XE, for a ground-truth caption \( c_t = \{w_t\}_{t=1}^T \), they ask the model to minimize the following cross-entropy loss:

\[
L_{xe} = -\sum_{t=1}^T \log P(w_t | w_{1:t-1}, I),
\]

where \( P(w_t | w_{1:t-1}, I) \) denotes the predicted probability of word \( w_t \).

When training with reinforcement learning, they firstly generate top-n captions \( \hat{C} = \{\hat{c}_i\}_{i=1}^n \) through beam search, and then optimize the following RL loss \([10]^2\):

\[
L_{rl} = -\frac{1}{n} \sum_{i=1}^n \left( r(\hat{c}_i, C) - b \right) \log p(\hat{c}_i),
\]

where \( r(\cdot, \cdot) \) is the reward function computed between \( \hat{c}_i \) and \( C \), and \( b = (\sum_{i=1}^n r(\hat{c}_i, C)) / n \) is the baseline, calculated as the mean of the rewards obtained by the generated captions.

4.2 TransDIC: Transformer-based Ref-DIC

Given a target image \( I_t \), we term all regions (or proposals) in it as target regions \( R_t = \{r_{t,i}\}_{i=1}^N \). Our proposed model tends to give each target region \( r_{t,i} \) some region references when generating captions. To this end, we firstly couple each \( r_{t,i} \) with semantic-similar regions from \( I_r \) as the reference, i.e., reference regions. Then we send each target region and its reference regions into the model for distinctive caption generation.

Specifically, our TransDIC consists of three components: 1) Similar Region Matching module in Section 4.2.1. 2) Two-Flow Encoder module. The module contains two parallel data flows to extract both target image and cross-image features in Section 4.2.2. 3) A plain Transformer-based Captioning Decoder for caption generation. An overview of our model is shown in Figure 4.

\[ \text{Figure 4: Overview of our proposed TransDIC model. It consists of three parts: (a) A Similar Region Matching module that couples each target region with reference regions. (b) A Two-Flow Encoder module that encodes both target and reference images. (c) A plain captioning decoder. In module (a), we use the same colors to denote the same objects. We firstly send target and reference images into the similar region matching module to construct Target-Reference tuples. Then we send constructed tuples into the Two-Flow Encoder for target image and cross-image features extraction. Finally, both kinds of features are sent into the decoder for caption generation.} \]

4.2.1 Similar Region Matching. For each target region, we retrieve regions from reference images with the highest similarity scores to it as its reference regions. Given a target image \( I_t \) and its corresponding \( I_r \), their region features are firstly projected into the memory space through an MLP layer. We denote memory features for \( I_t \) and the K images in \( I_r \) as \( M_t = \{m^t_k\}_{k=1}^N \) and \( M_k = \{m^k_i\}_{i=1}^M \), respectively.

Then, we calculate the cosine similarity scores between features in \( M_t \) and \( M_k \), i.e.,

\[
S_{ij}^K = \cos(m^t_k, m^k_j),
\]

where \( m^t_k \) represents the \( j \)-th region in \( I_t \), and \( m^k_i \) represents the \( i \)-th region in reference image \( I_k \). We apply max operation to get the most similar region for \( r^t_j \) according to the calculated \( S^K \):

\[
j^t_j = \arg \max_i (S_{ij}^K)_{i=1}^N.
\]

Where \( j^t_k \) denotes the most similar region from image \( I_k \) for \( r^t_j \), i.e., reference region. As an example in Figure 4 (a), reference image \( I_k \) contains regions “couch”, “painting” and “TV”. For the region “couch” in \( I_t \), we can learn from the similarity matrix \( S^K \): the region is similar to the “couch” in \( I_K \) (deep green) while it is different from the “TV” or “painting” (light green) in \( I_k \). Max operation is then token along each column of \( S^K \), and “couch” in \( I_k \) is selected as the reference region for “couch” in \( I_t \).

Finally, for each target region \( r^t_n \), we gather \( K \) reference regions, one for each, from \( K \) reference images. We put these \( K + 1 \) regions together and term them as a Target-Reference region tuple:

\[
T_n = \{r^t_n, r^{t_1}_n, r^{t_2}_n, \ldots, r^{t_K}_n\}. \quad n = \{1, \ldots, N\}
\]

For example in Figure 4 (a), all regions marked with green boxes form a Target-Reference tuple (for “couch”).
4.2.2 Two-Flow Encoder. Our proposed module takes $M_t$ and $M_b$ as input, and extracts target image features and cross-image features through the Target flow and the Target-Reference flow, respectively. An overview of the module is shown in Figure 4 (b).

**Target flow.** The flow enables the region interactions within the target image $I_t$. Same as the standard transformer-based captioning model, it sends memory features $M_t$ into multiple self-attention layers, and finally outputs encoded features $\hat{M}_t = \{\hat{m}_t^i\}_{i=1}^N$ for $I_t$.

**Target-Reference flow.** This data flow consists of select layers and fuse layers. Given a Target-Reference region tuple $T_n$, we denote the memory features for target region and reference regions in $T_n$ as $\hat{m}_n^t$ and $\hat{m}_n^r$ respectively. The flow takes in those two kinds of features and generates cross-image features through select and fuse layers:

**Fuse layer.** The goal of the fuse layer is to enable the interactions among memory features within $M_n$. We stack multiple fuse layers, and the $l$-th fuse layer is calculated as follows:

$$U_l = \text{MH}(U_{l-1}, U_{l-1}, U_{l-1}),$$

where $U_0$ refers to $\hat{M}_n$ and $U_{l-1}$ is the outputs of the $(l-1)$-th fuse layer. Because all the features in $\hat{M}_n$ are semantic-similar, the model can learn to capture the primary concepts they are describing.

**Select layer.** The select layer builds on the co-attention mechanism. We set features of target region as query, features of reference region as key and value in multi-head attention. Multiple mechanisms. We set features of target region as query, features of reference region as key and value in multi-head attention. Multiple

$$V_l = \text{MH}(V_{l-1}, U_{l-1}, U_{l-1}),$$

where $V_0$ refers to $\hat{m}_n^t$, $V_{l-1}$ and $U_{l-1}$ are the outputs of the $(l-1)$-th select and fuse layer, respectively. By the residual connection in self-attention blocks, feature $\hat{m}_n^t$ will gradually select useful information from reference images while preserving the original information from $I_t$.

As an example shown in Figure 4 (b), our model can learn to focus on the unique attributes and objects in $I_t$. For unique attributes, we send all reference regions of “couch” (green boxes) into fuse layers, the model will be informed they are describing the concept “couch”. Since target region also describes “couch”, the select layer learns to focus on the unique color “red” of the “couch” in $I_t$. When predicting unique objects, for “vase” region in $I_t$, because all selected reference regions for it (blue boxes) do not contain the same concept, the select layer learns that “vase” is a unique object in $I_t$.

We use the outputs of the last select layer as final refined target feature $\hat{m}_n^t$ for region $r_n^t$ in $I_t$. For N Target-Reference tuples in $I_t$, we can get $\hat{M}_t = \{\hat{m}_t^i\}_{i=1}^N$ as the outputs of Target-Reference flow. Finally, we concatenate the outputs of Target flow and Target-Reference flow as the final outputs of the Two-Flow Encoder:

$$M'_t = [\hat{M}_t; \hat{M}_t],$$

where $[\cdot; \cdot]$ denotes concatenation operation, and $M'_t$ will be sent into decoder for caption generation.

5 EXPERIMENTS

In this section, we describe the datasets used for experiments and introduce a new distinctiveness-based evaluation metric DisCIDEr. We conduct extensive experiments and ablation studies to reveal the

superiority of our proposed model, and our proposed benchmarks for Ref-DIC.

5.1 Datasets

We developed the COCO-DIC and Flickr30K-DIC based on the MS-COCO [9] and Flickr30K [29]. They contain 123287 and 31014 images, respectively. Each image is annotated with 5 ground-truth captions. For both datasets, we followed the splits provided by [16], and constructed reference image groups within the training, validation, and test splits. For completeness, we also reported results on Wang et al. [43]’s dataset for Ref-DIC.

5.2 Evaluation Metrics

We applied two kinds of metrics to evaluate the accuracy and distinctiveness of generated captions. For accuracy evaluation, we calculated four commonly used evaluation metrics: BLEU-N (B-N) (1- to 4-grams) [27], ROUGE-L (R) [20], METEOR (M) [3], and CIDEr (C) [40]. For distinctiveness evaluation, we developed a new metric named DisCIDEr (DisC). We introduce the DisCIDEr below. DisCIDEr. All existing metrics designed for Ref-DIC task fail to fully explore the distinctiveness of each individual n-gram in GT captions of target image. To solve this, we assign these n-grams with different weights according to group-level distinctiveness: if an n-gram occurs frequently in ground-truth captions of reference images, it is less distinctive. As an example in Figure 5, both target and reference images are describing “fireplace”, so we should assign lower weights to the word “red” in ground-truth captions of $I_t$ at evaluation time. Instead, since only image $I_t$ contains the object “fireplace”, we should put more weight on it.

To realize this intuition, we modify the n-gram weighting procedure of CIDEr by adding a re-weight term. For a target image $I_t$, we denote its generated and ground-truth captions as $c$ and $s_t = \{s'_t\}_{j=1}^M$, respectively. Similarly, we denote ground-truth captions for reference images $I_r$ as $S_r = \{s'_r\}_{j=1}^M$, $r = \{1 \ldots K\}$. Then the number of times an n-gram $s'_t$ occurs in $s'_r$ is denoted as $h_d(s'_t)$. We modify CIDEr by adding an inverse reference frequency term after it when calculating $g_d(s'_t)$ for the n-grams in ground-truth

Figure 5: An example of the intuition behind DisCIDEr. N-gram “red” appears in both target and reference images, thus should be given less attention (blue). In contrast, we should pay more attention to “fireplace”, because it appears only in target image (red).
Table 2: Comparison of captions accuracy on COCO family with state-of-the-art image captioning models.

| Model  | B-1  | B-4  | M   | R   | C   | DisC |
|--------|------|------|-----|-----|-----|------|
| Dataset: MS-COCO |
| UpDown [1] | 79.8 | 36.3 | 27.7 | 56.9 | 120.1 | — |
| AoANet [14] | 80.2 | 38.9 | 29.2 | 58.8 | 129.8 | — |
| Transformer [38] | 80.0 | 38.2 | 28.9 | 58.2 | 127.3 | 98.7 |
| M$^2$ Transformer [10] | 80.8 | 39.1 | 29.2 | 58.6 | 131.2 | — |
| GdisCap [43] | 80.0 | 37.3 | 28.4 | 57.5 | 125.8 | 96.6 |
| CAGC [19] | 80.7 | 38.1 | 28.7 | 57.9 | 127.9 | 98.0 |
| TransDIC (Ours) | 81.6 | 39.3 | 29.2 | 58.5 | 132.0 | 102.2 |

| Dataset: COCO-DIC |
| GdisCap [43] | 80.0 | 37.3 | 28.4 | 57.5 | 125.8 | 96.6 |
| CAGC [19] | 80.7 | 38.1 | 28.7 | 57.9 | 127.9 | 98.0 |
| TransDIC (Ours) | 81.0 | 38.8 | 29.1 | 58.2 | 130.8 | 101.9 |

Table 3: Comparison of captions accuracy on Flickr30K family with state-of-the-art image captioning models.

| Model  | B-1  | B-4  | M   | R   | C   | DisC |
|--------|------|------|-----|-----|-----|------|
| Dataset: Flickr30K |
| NIC [41] | 66.3 | 18.3 | —   | —   | —   | —   |
| Xu et al. [46] | 66.9 | 19.9 | 18.5 | —   | —   | —   |
| Transformer [38] | 70.7 | 27.7 | 21.4 | 49.0 | 61.2 | 39.1 |
| GdisCap [43] | 71.7 | 29.0 | 22.1 | 49.6 | 65.6 | 41.2 |
| CAGC [19] | 72.9 | 29.1 | 21.9 | 50.1 | 62.2 | 39.0 |
| TransDIC | 73.2 | 30.1 | 22.5 | 50.3 | 65.1 | 41.4 |

Table 4: Ablation study of Two-Flow Encoder on COCO-DIC. “Fuse” and “Select” denote the fuse layer and select layer in the Target-Reference flow, respectively.

| Fuse | Select | B-1  | B-4  | M   | R   | C   | DisC |
|------|--------|------|------|-----|-----|-----|------|
| X    | X      | 80.0 | 38.2 | 28.9 | 58.2 | 127.3 | 98.7 |
| ✓    | ✓      | 81.3 | 38.4 | 29.0 | 58.1 | 130.0 | 100.2 |
| ✓    | X      | 81.6 | 39.1 | 29.1 | 58.4 | 131.3 | 101.6 |
| ✓    | ✓      | 81.6 | 39.3 | 29.1 | 58.5 | 132.0 | 102.2 |

Results on COCO-family Benchmarks. From Table 2, we can observe: 1) For accuracy evaluation, our proposed TransDIC achieves the best performance on all conventional metrics at both COCO-DIC and Wang et al. [43] (e.g., 132.0 vs. 125.8 in GdisCap on CIDEr). Meanwhile, our model outperforms some strong state-of-the-art models (e.g., 132.0 vs. 131.2 in $M^2$ Transformer on CIDEr) in terms of accuracy-based metrics. 2) For distinctiveness evaluation, our model gets the highest scores on DisCIDEr in the two datasets.

Results on Flickr30K-family Benchmarks. From Table 3, we can observe: 1) For accuracy evaluation, our TransDIC achieves the largest performance gains on most metrics while it is narrowly defeated by GdisCap on CIDEr (65.1 vs. 65.6). 2) For distinctiveness-based metrics, our model outperforms GdisCap by 0.2 on DisCIDEr (41.4 vs. 41.2), despite it having a lower CIDEr score.

5.5 Ablation Studies

We conducted extensive experiments to verify the influences of the proposed Two-Flow Encoder module and group similarity.

5.5.1 Influence of Two-Flow Encoder. To measure the influence of each component in our proposed Target-Reference flow, we trained an ordinary transformer as the baseline and three variants of our model. 1) Target-Reference flow only contains the select layer: stacked select layers always take the original reference features as input. 2) Target-Reference flow only contains the fuse layer: outputs of the last fuse layer are directly used as the outputs of the Target-Reference flow. 3) Transformer with complete select and fuse layers. All these models were trained on COCO-DIC and the results were shown in Table 4. Results. As can be observed in rows 2 and 3, two additional components can improve captioning performance consistently in terms of both accuracy and distinctiveness. Above all, our complete model achieves the most promising performance in all metrics.
5.5.2 Influence of Group Similarity. To quantify the influence of group similarity, we used different $p$ when choosing $K$ most similar images from $I_f$, e.g., top-2 to top-6 (top2-6) group when setting $p$ to 2. The results were reported in Table 5.

Results. From Table 5 (a), we can observe: Top1-5 group is surpassed by top3-7 group on DisCIDEr (101.4 vs. 102.2), despite it having a marginal improvement on CIDEr (132.2 vs. 132.0). The results indicate that the most similar reference group is not always helpful to group-level distinctiveness.

5.6 Qualitative Results

We illustrated the qualitative results of our proposed TransDIC and compared it with the Transformer and SOTA Ref-DIC model GdisCap [43] in Figure 6. Naive captioning models generate identical captions for similar images. In contrast, our TransDIC can describe the unique attributes and objects in the target image. For unique attributes, as shown in Figure 6 (left), TransDIC precisely captures the unique attributes “green and white” and “yellow and red” for the two trains, respectively. In Figure 6 (right), for unique objects, TransDIC captures unique objects “carrots” and “a fork” for each individual image. The results demonstrate that TransDIC can generate distinctive captions in terms of unique objects and attributes.

6 LIMITATIONS

One possible limitation of our work is that if original human-annotated captions omit some objects or attributes, it will lead to: 1) our proposed two-stage matching mechanism may fail to collect object-attribute level similarity reference images. 2) our proposed DisCIDEr may degrade to existing CIDEr. We believe this omitting problem is due to the natural defect of datasets (COCO/Flickr30K): Since these datasets are annotated for general captioning task, human annotators may tend to simply describe the objects while ignoring its attributes (e.g., color) when there is no reference image.

7 CONCLUSIONS

In this paper, we argued that all existing DIC works fail to achieve group-level distinctiveness. To solve this problem, firstly, we introduced a two-stage matching mechanism and proposed two new benchmarks for Ref-DIC. Then, we developed a Transformer-based model for Ref-DIC. Finally, we came up with a new evaluation metric termed DisCIDEr. We conducted extensive experiments to verify the effectiveness of TransDIC. Moving forward, we are going to 1) extend our Ref-DIC into video domains, or 2) design stronger Ref-DIC models.

ACKNOWLEDGMENTS

This work was supported by the National Key Research & Development Project of China (2021ZD0110700), the National Natural Science Foundation of China (U19B2043, 61976185), Zhejiang Natural Science Foundation (LR19F020002), Zhejiang Innovation Foundation(2019R52002), and the Fundamental Research Funds for the Central Universities(226-2022-00087).
Appendix. A USER STUDY INTERFACE

We conduct user studies to validate the effectiveness of our proposed DisCIDEr. Specifically, we invite 5 experts and give them a picture and two captions that describe it. They are asked to choose which one better matches the image in terms of accuracy and distinctiveness and the user interface for this is shown in Fig. 7. The caption which got more than 3 votes is regarded as human judgment. We randomly select 100 images from the test split of MS-COCO and ask the experts to give their judgments. A metric agrees with human judgment only if it gives a higher score to the same caption that experts choose. We use the agreement counts between human judgment and the metric as the effectiveness measurement.

Following is an image and two captions that describe it:

A group boats are setting in the water.

A group boats are docked in a marina.

Please chose which caption better describe the image in terms of accuracy and distinctiveness.

Figure 7: We display an image and two captions generated by Transformer [38] and our proposed TransDIC, respectively. The users are asked to choose which caption better describe the image in terms of accuracy and distinctiveness.

User Studies on DisCIDEr. We conducted a user study to validate the effectiveness of DisCIDEr with 5 experts. We randomly selected 100 images (100 trials) from the test set. In each trial, an image and two captions (generated by Transformer and TransDIC, respectively) were displayed and these experts are asked to choose the better caption in terms of distinctiveness and accuracy. These captions which got more than 3 votes were regarded as human judgment. We calculated the agreements between human judgments and different metrics (i.e., whether humans and metrics give higher scores to a same caption). Results were reported in Table 6. From Table 6, we can observe that DisCIDEr achieves better agreement than both the accuracy-based metric CIDEr and the distinctiveness-based metric DisWordRate.

Table 6: User study of different metrics

| Metric          | CIDEr [40] | DisWordRate [43] | DisCIDEr |
|-----------------|------------|-------------------|----------|
| Agreements      | 58         | 62                | 64       |

Appendix. B DETAILS OF THESE COMPARED BASELINES

In this section, we describe implementation details of two compared state-of-the-art baselines: GdisCap [43] and CAGC [19]. These two group-based captioning baselines are both built on top of the widely-used Transformer [38] architecture, which takes a sequence of visual features as input and encode them through multiple self-attention layers. We introduce the two baselines in detail.

1. **GdisCap [43]:** It develops a Group-based Memory Attention (GMA) module which assigns higher weights to the distinctive regions and two special losses: DisLoss and MemCls loss, which directly encourage the model to generate distinctive words. We re-implement their proposed GMA module and DisLoss.

2. **CAGC [19]:** It is designed to describe a group of target images in the context of another group of related reference images. To this end, they extract visual features for all images in target and reference groups from the ResNet50 network [13] (after pool5 layer). Then, these Visual features of images are sent into multiple self-attention layers to enable the interactions within each group (target image interacts with other images only within target group, and same for reference group). Finally, they construct group representations and contrastive representations for both groups and send them into a decoder for captions generation. To make a fair comparison, we replace the visual features with mean pooled bottom-up features [1] and re-implement their visual content encoder.
Appendix. C MORE QUALITATIVE RESULTS
We report more qualitative results in Fig. 8 and Fig. 9. In Fig. 8, we compare our TransDIC with Transformer and GdisCap. Remarkably, our model can precisely capture the unique objects and attributes in the image. For example, in Fig. 8 (row 2, column 3), both Transformer and GdisCap wrongly mix the tie and shirt into “blue shirt” while TransDIC correctly matches the “shirt” with “black” and “tie” with “blue”. In terms of numerals, in Fig. 8 (row 2, column 1), only TransDIC predicts the right number, whereas others mistake it for “one”. In Fig. 9, we test all models with two similar images, as can be seen on the left side, TransDIC correctly captures the “red and white” and “green and red” for the two buses, respectively.

Figure 8: Examples of generated captions for two similar images using Transformer, GdisCap and TransDIC. The green words indicate the unique details in the images while blue denote the mistakes in the generated captions.

Figure 9: Examples of generated captions for two similar images. The green words indicate the unique details in the images.