Technical Report of NICE Challenge at CVPR 2024: Caption Re-ranking Evaluation Using Ensembled CLIP and Consensus Scores

Kiyoon Jeong*, Woojun Lee*, Woongchan Nam, Minjeong Ma, Pilsung Kang†
School of Industrial and Management Engineering, Korea University
{kiyoon_jeong, woojun_lee, woongchan_nam, minjeong_ma, pilsung_kang}@korea.ac.kr

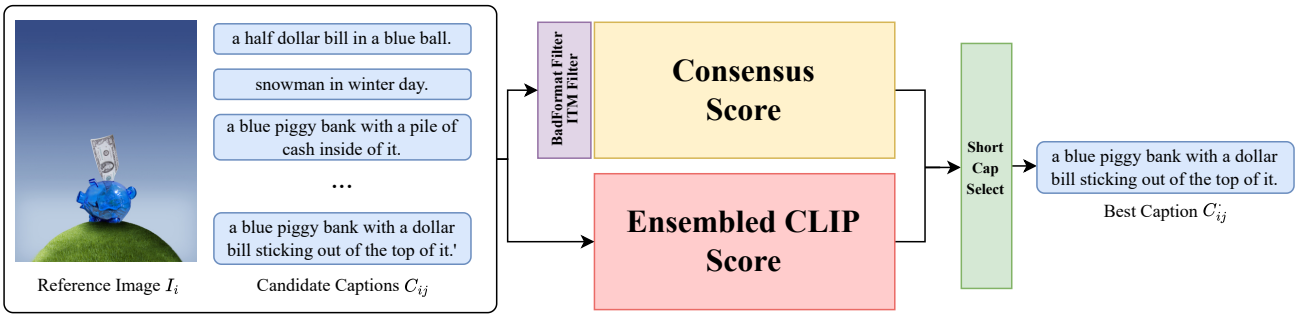

Figure 1. Overview of **ECO** (**E**nsembled **C**lip score and c**O**nsensus score) framework utilizes both the Ensembled CLIP score and Consensus score to select the ideal caption.

Abstract

*This report presents the **ECO** (**E**nsembled **C**lip score and c**O**nsensus score) pipeline from team **DSBA LAB**, which is a new framework used to evaluate and rank captions for a given image. **ECO** selects the most accurate caption describing image. It is made possible by combining an Ensembled CLIP score, which considers the semantic alignment between the image and captions, with a Consensus score that accounts for the essentialness of the captions. Using this framework, we achieved notable success in the CVPR 2024 Workshop Challenge on Caption Re-ranking Evaluation at the New Frontiers for Zero-Shot Image Captioning Evaluation (NICE). Specifically, we secured third place based on the CIDEr metric, second in both the SPICE and METEOR metrics, and first in the ROUGE-L and all BLEU Score metrics. The code and configuration for the **ECO** framework are available at* https://github.com/DSBA-Lab/ECO.

1. Introduction

The NICE 2024 Challenge Caption Re-ranking track is a competition that challenges participants to identify the most accurate and comprehensive caption from a set of candidate captions for a given image. The goal is to select caption that accurately and thoroughly describe an image. The NICE dataset provided for this challenge comprises 20,000 images, and about 60 captions for each image, forming a zero-shot evaluation dataset. It includes various images, candidate captions generated by different models, and undisclosed answer captions created by human annotators.

Participants in the 'Image Caption Re-ranking' task should choose and submit the caption they consider most appropriate for each image. Their submissions are evaluated against five undisclosed correct captions written by different human annotators, based on five metrics: CIDEr [13], SPICE [1], METEOR[2], ROGUE-L[6], and BLEU[8]. The aim is to encourage innovative approaches to selecting captions that enhance the accuracy and depth of image descriptions.

Upon first glance, one may assume this task is similar to the image-to-text retrieval task. However, there is a distinction as the Retrieval Task involves selecting the clear answer caption among multiple candidate captions, some of

*Equal Contribution.
†Corresponding author.

which may significantly deviate from the answer. This task is evaluated using recall metrics such as recall@1, 5, and 10 [10]. On the other hand, 'Image Caption Re-ranking' task literally focuses on re-ranking candidate captions, which are closely aligned with the correct answers. These differences call for evaluating captions in more detail.

To develop an algorithm capable of re-ranking captions based on the quality of their descriptions of images, it was important first to establish a clear definition of what constitutes an "accurate and thorough" caption. To achieve this, we determined that an ideal caption must meet two key criteria:

1. An ideal caption should have a high semantic alignment with the associated image.
2. An ideal caption should have a high degree of essentialness.

The first criterion for a caption is that it should accurately reflect the context of the image and not include any content that is not present in the image. In other words, the caption should be semantically aligned with the image, and the more alignment there is, the better it will meet the first criterion.

The second criterion requires that the caption avoids using overly elaborate language and instead focuses on using essential expressions. This means that the caption should only include expressions that are necessary to describe the image. The more indispensable each expression is for accurately depicting the image, the better it meets the second criterion.

It is important to note that meeting only one of the criteria does not guarantee the caption is ideal. A caption with high semantic alignment might still have non-essential elements, while focusing solely on essential elements might not adequately represent the image. To address this issue, we propose the **ECO** framework, which uses scoring methods to evaluate both the degree of semantic alignment between the image and caption, and how essential the terms in the captions are.

To determine how well captions match images, we used various pre-trained CLIP [9] models and BLIP-2[5] model to calculate the cosine similarity between image and text features. We then created a robust Ensembled CLIP score by combining the results. To measure how essential the terms in the captions are, we used a Consensus score derived from comparing candidates within the pool of captions. Finally, we combined the Ensembled CLIP score and Consensus score to calculate the final score. If the difference between the top two captions for an image is negligible, both captions are considered equally good at describing the image. In this case, we choose the caption with fewer words as the final caption.

The proposed **ECO** framework is a method for caption evaluation that is easy to understand and doesn't require any additional fine-tuning. It can take into account both the alignment of images and text, as well as the essentialness of captions in a zero-shot setting. The overall framework can be seen in Fig. 1. By using this approach, we were able to achieve impressive results in the NICE 2024 Challenge. We came in third place based on the CIDEr metric, second place in both SPICE and METEOR metrics, and first place in the ROUGE-L and BLEU metrics. These achievements show that our method is not limited to excelling in a single metric, but is versatile and can be applied to various evaluation criteria.

2. Proposed Method

Building on the concept of a 'well-explained image caption' defined in Sec. 1, we introduce **ECO**, a framework designed to select the ideal caption by considering both the semantic alignment between images and captions and the essentialness of captions. **ECO** comprises two main scoring algorithms: 1) the Ensembled CLIP score and 2) the Consensus score. In Sec. 2.1 discusses the integrated CLIP score derived from the cosine similarity between image and caption embeddings, utilizing a variety of CLIP models and the ITC Loss calculation from BLIP-2 to assess the alignment between images and captions. In Sec. 2.2 covers the method of measuring essentialness through mutual comparison between candidate captions, termed the Consensus score. In Sec. 3.1, we detail the process of integrating these two scores, and Sec. 2.5 explains how the combined score is used to select the final caption.

2.1. Ensembled CLIP score

The CLIP score[4] is a metric that measures the semantic alignment between images and captions by comparing the cosine similarity between the image embedding $\mathbf{E}_I$ and caption embedding $\mathbf{E}_C$ through a pre-trained CLIP model. However, as the training data for the pre-trained CLIP model differs from the zero-shot caption re-ranking dataset provided, the accuracy of the single CLIP score may not be reliable. To address this issue, an ensemble of CLIP scores from various models that have proven to perform well in zero-shot tasks can provide a more robust semantic alignment than a single CLIP score.

$$S_{\text{CLIP}}(I, C) = \cos(E_I, E_C) \tag{1}$$

$$S' = \frac{S - \text{mean}(S)}{\text{std}(S)}. \tag{2}$$

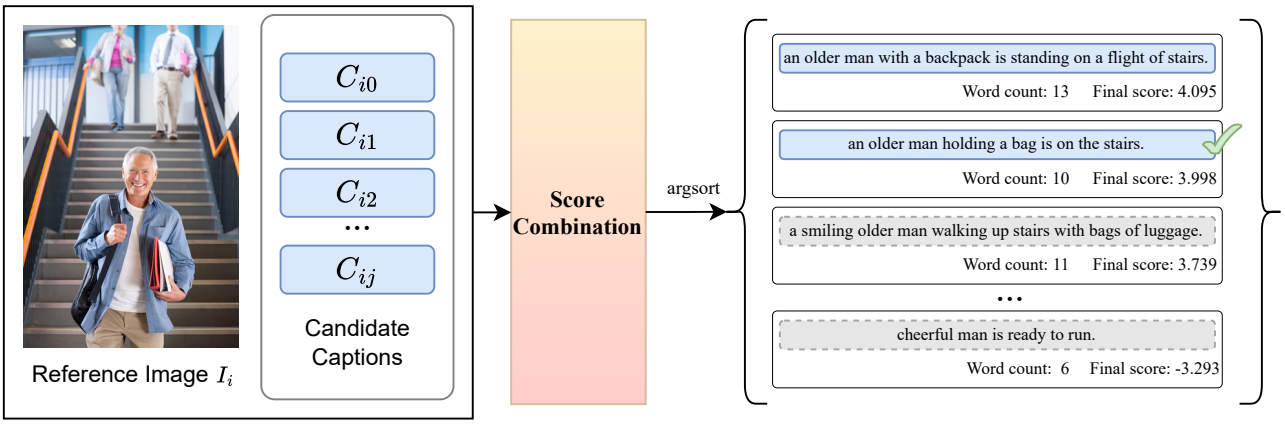

Figure 2. The Short Cap Selection process: If the final scores of the top 2 captions differ by less than the threshold θ, a caption with fewer words is chosen.

$$S_{\text{ensemble}} = \sum_{i \in I} {S^i_{\text{CLIP}}}', \quad \forall i \in I,$$
$$\text{where } I = \{\text{“EVA-CLIP”, “MetaCLIP”, “MobileCLIP”, “OpenCLIP”, “BLIP-2”}\}. \tag{3}$$

The conventional CLIP score determines semantic alignment using the cosine similarity between $\mathbf{E}_I$ and $\mathbf{E}_C$, substituting any negative values with zero. However, the **ECO** framework allows negative values to be remained to achieve a more refined score distribution for image-caption pairs with low relevance. To calculate the Ensembled CLIP score, the cosine similarity values between $\mathbf{E}_I$ and $\mathbf{E}_C$ were calculated using models such as EVA-CLIP-18B[1] [11], MetaCLIP[2] [14], MobileCLIP[3] [12], OpenCLIP[4] [3], and BLIP-2[5].

2.2. Consensus score

We refer to the extent to which a caption is made up of essential expressions as its ”Essentialness”. When various models produce different captions, the expressions that appear most often are considered essential to describe the image. To measure this essentialness, we use a scoring method called the Consensus score.

The Consensus score is a metric derived from the CIDEr score, that calculates the TF-IDF weights for N-Grams across candidate and reference captions. It then calculates the cosine similarity between the TF-IDF weight vectors of the candidate caption and each reference caption. To assess the essentialness of expressions within a caption, we calculate the Consensus score for each caption, using all remaining candidate captions as reference captions, except the caption under evaluation.

However, the effectiveness of the Consensus score is significantly influenced by the quality of the caption pool used as references. In other words, if the reference caption set consists only of high-quality captions, the Consensus score is more likely to reliably reflect the degree of essentialness. To enhance the effectiveness of the Consensus score, we use two filters to make a high-quality caption pool.

2.3. Caption Filtering

The consensus scoring system gives higher scores to captions that use essential words frequently used in multiple captions. However, if the pool of candidate captions has many irrelevant or non-conforming captions, this scoring method may not work. To solve this issue, we filter the candidate caption pool with two types of filtering.

2.3.1 Bad Format Filter

Based on the insights from the Flickr30k [15] and COCO [7] datasets, we have identified the typical structure of captions. Generally, a caption is a phrase or clause of a single sentence and includes a sufficient amount of information in an image. To ensure high-quality and relevant captions, we filtered out captions that contained more than two periods or more than three commas, or those that had fewer than five words. This filtering was done systematically using a rule-based algorithm. This process helps to ensure that the captions being evaluated adhere to conventional standards.

[1] https://github.com/baaivision/EVA/tree/master/EVA-CLIP-18B
[2] https://github.com/facebookresearch/MetaCLIP
[3] https://github.com/apple/ml-mobileclip
[4] https://github.com/mlfoundations/open_clip
[5] https://github.com/salesforce/LAVIS/tree/main/projects/blip2

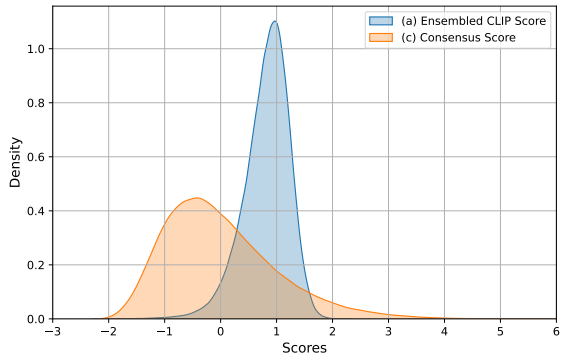

Figure 3. Comparison of the Ensembled CLIP score and Consensus score distributions

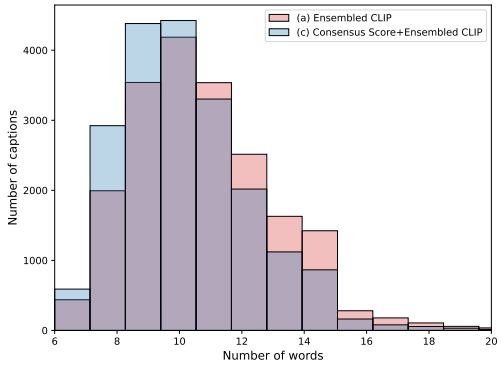

Figure 4. Comparison of the number of words in captions selected by the Ensembled CLIP score versus the Consensus score.

2.3.2 ITM Filter

In order to filter out captions that are irrelevant to the content of the image from group of candidates, we implement another filter which is called ITM filter. Removing captions that are not related to the image is important because they can hinder the consensus scoring. The consensus scoring is based on the agreement among captions and can be affected by the inclusion of expressions that are not related to the image content.

To filter out irrelevant captions, the Image-Text Matching (ITM) Loss from BLIP-2 is used. The ITM loss is designed to classify an image and text pair as either positive or negative, making it very efficient for filtering out captions that are not related to the image.

The ITM loss is calculated for each caption associated with an image, and the top 50% of captions with the highest ITM values are selected. These captions are then used in the caption pool for consensus scoring, ensuring that the captions considered are more likely to be relevant and aligned with the image content.

2.4. Score Combination

In Sec. 2.1 and Sec. 2.2, we defined the Ensembled CLIP score and the Consensus score. After normalizing these scores individually, we combine them using a weighted sum to form the final score. This approach allows us to adjust the influence of each score differently, ensuring that both the semantic alignment between the image and captions and the essentialness of the captions are appropriately considered in determining the most suitable caption. This method of integration provides a flexible framework that can be tailored to prioritize different aspects of caption quality depending on the specific requirements of the task at hand.

$$S_{comb} = \lambda_1 S'_{ensemble} + \lambda_2 S'_{consensus}. \quad (4)$$

2.5. Short Caption Selection

By combining the earlier Ensembled CLIP score and Consensus score, we obtained a final score that reflects both the semantic alignment between the caption and image, and the essentialness of the caption. If there is a clear distinction in the final score, the caption with the highest score is selected as the optimal caption. In cases where the difference in the final score is not pronounced, meaning the difference between the scores of the top-2 captions is less than a threshold θ, we chose the caption with fewer words as the final caption from the perspective of essentialness as shown in Fig. 2.

3. Experiments

3.1. Score Combination Setting

When setting the weights for score combination, we observed significant differences in outcomes depending on how the Consensus score and the Ensembled CLIP score score were utilized. When comparing selected captions by using only the Consensus score to those by using Consensus scoreand Ensembled CLIP score equally, We find a difference in 5,396 out of 20,000 captions. Conversely, the discrepancy reached 18,217 captions when the Ensembled CLIP score was used alone versus when it was combined with the Consensus score. To analyze these differences accurately, we visualize the distribution of both scores after normalization and discovered that the maximum value of the Consensus score was approximately three times larger than that of the clip score as shown in Fig. 3. This discrepancy suggested that, in situations where the caption with the highest combined score was selected, the overwhelming influence of the Consensus score could skew the results.

To ensure a balanced reflection of both scores, we decided to set λ_2 (the weight for the Consensus score) larger than λ_1 (the weight for the Ensembled CLIP score). After a few experiments, we confirm that a ratio of 1:3.52 is

Method	CIDEr	SPICE	METEOR	ROUGE-L	BLEU_{AVG}
(a) Ensembled CLIP Score	176.83	30.06	34.84	63.88	57.88
(b) Consensus Score	202.89	31.21	35.79	68.20	67.07
(c) Ensembled CLIP Score + Consensus Score	212.44	32.41	36.98	69.41	**67.90**
(d) Ensembled CLIP Score + 3.52× Consensus Score	**218.47**	**33.46**	**37.98**	**70.13**	67.58

Table 1. **Ablations of Score Combination**: The caption with the highest score is ultimately selected for submission. This involves **(a)** combining the CLIP score from models like EVA-CLIP-18b, MetaCLIP, MobileCLIP, and OpenCLIP with the BLIP-2 ITC score, **(b)** using consensus-based scoring alongside Caption Filtering, **(c)** combining the Ensembled CLIP score with the Consensus score at a 1:1 ratio, and **(d)** combining the Ensembled CLIP score with the Consensus score at a ratio of 1:3.52.

Method	CIDEr	SPICE	METEOR	ROUGE-L	BLEU_{AVG}
(b) Consensus Score	**202.89**	**31.21**	**35.79**	**68.20**	**67.07**
(e) Consensus Score (w/o Caption Filtering)	192.98	30.10	34.41	65.89	64.75

Table 2. **Ablations of Caption Filtering. (e)** has the same settings as **(b)** with the exception of using a caption filter.

Method	CIDEr	SPICE	METEOR	ROUGE-L	BLEU_{AVG}
(d) Ensembled CLIP Score + 3.52*Consensus Score	218.47	**33.46**	**37.98**	70.13	67.58
(f) + Short Capation Selection	**220.53**	33.01	37.68	**70.31**	**69.20**

Table 3. **Ablations of Short Caption Selection. (f)** has the same settings as **(d)** with the addition of Short Caption Selection. We set the threshold θ to 0.39.

the most effective. This decision is supported by experimental evidence presented in Tab. 1, where the CIDEr score for results combined equally is 212.44, compared to 218.46 for combinations using the 1:3.52 ratio. This result confirm that placing greater weight on the λ_2 leads to improved outcomes. This weighting strategy aims to balance the influence of both the Consensus score and the Ensembled CLIP score, ensuring that both semantic alignment and essentialness are appropriately considered in the final caption selection.

3.2. Consensus Scoring's Effectiveness in Identifying Essentialness

We conduct an evaluation of the effectiveness of using the ITM Filter and Bad Format Filter, by comparing Consensus score for captions with and without filtering. Based on the results presented in Tab. 2, we find that the filtered case has a Consensus score of 202.89, while the unfiltered case has a score of 192.98. This indicates that filtering the caption pool improves the quality of captions selected, as measured by the CIDEr metric.

Furthermore, we created a visualization in Fig. 4 to show the length of captions selected from each pool, measured by the number of words per caption. The visualization shows that captions chosen from the filtered pool are significantly shorter on average within their respective pools. These findings collectively suggest that filtering the caption pool enhances the ability of consensus scoring to assess essentialness. By improving the overall quality of captions selected, this strategy maximizes the functionality of consensus scoring.

3.3. Effects of Caption Filtering

To assess the effectiveness of the ITM Filter and Bad Format Filter, we compare the evaluation results of the filtered cases with those of the unfiltered cases. The results, Tab. 2, shows that the CIDEr score increases from 192.98 to 202.89 and there are also improvements in every other metrics. It demonstrate an enhancement in consensus scoring by refining the pool of captions. Furthermore, we visualized the relative rank of the selected caption within the candidate caption pool in terms of the number of words. Fig. 5 reveals that, after filtering the pool, the chosen captions are significantly shorter than those in their respective pools. These findings collectively suggest that filtering the caption pool enhances the ability of consensus scoring to discern essentialness, maximizing its effectiveness in evaluating captions.

3.4. Effects of the Short Caption Selection

When comparing the results of applying Short Caption Selection to those without it, as shown in Tab. 3, it's evident that using Short Caption Selection improves performance: the CIDEr score increased from 218.47 to 220.53 upon ap-

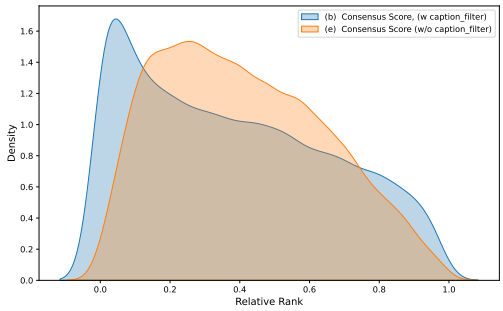

Figure 5. The relative rank of the selected caption within the candidate caption pool in terms of the number of words.

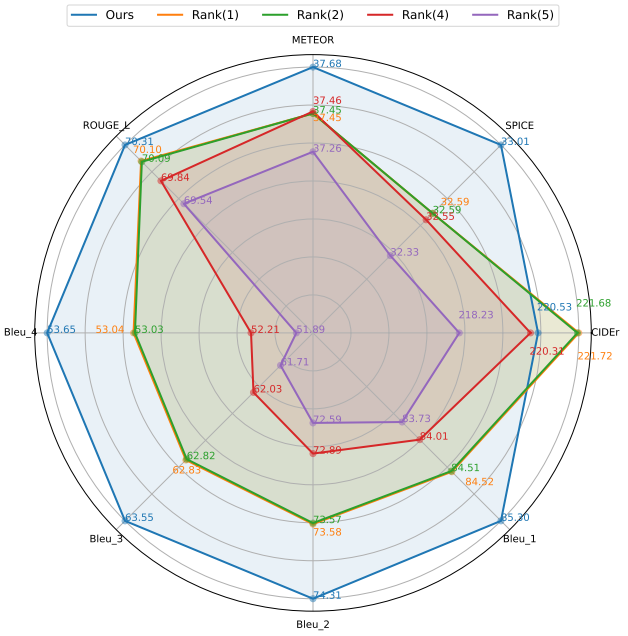

Figure 6. Comparative advantage of our methodology. Our approach demonstrates a marked improvement over other top-ranked methodologies.

plication. Also there are improvement in some other metrics (ROUGE-L, BLEU) when Short Caption Selection was applied, compared to when it was not. This shows that choosing shorter captions can make the caption selection process more effective.

4. Conclusion

We propose **ECO**, a zero-shot caption re-ranking framework that incorporates both image-caption semantic alignment and caption essentialness. Our method selects the most ideal caption for an image from several candidates, without model training. Through Fig. 6, we have verified that our methodology serves as a general caption re-ranking framework that performs well across all metrics, demonstrating its effectiveness and versatility in identifying ideal captions.

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer, 2016. 1

[2] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 1

[3] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 3

[4] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022. 2

[5] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2

[6] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 1

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 3

[8] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 1

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[10] Gerard Salton. *The SMART retrieval system—experiments in automatic document processing*. Prentice-Hall, Inc., 1971. 2

[11] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 3

[12] Pavan Kumar Anasosalu Vasu, Hadi Pouransari, Fartash Faghri, Raviteja Vemulapalli, and Oncel Tuzel. Mobileclip: Fast image-text models through multi-modal reinforced training. *arXiv preprint arXiv:2311.17049*, 2023. 3

[13] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 1

[14] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023. 3

[15] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 3