A Concept-Guided Decoding Mechanism for Diagnostic Captioning

Anonymous ACL submission

Abstract

The automated generation of a diagnostic text based on a set of radiology images is known as Diagnostic Captioning. This draft outlines the medical conditions of a patient, assisting the clinician by providing an initial estimation of the patient's condition, speeding up and safeguarding the diagnostic process. The accuracy of a diagnostic text, however, depends strongly on how effectively the key medical concepts depicted in the respective radiology image are represented. A prevalent strategy used to steer a Language Model (LM) towards a desired behaviour and the prevention of inaccurate content generation is the employment of task-specific guided decoding methods. In this work, we propose a novel guided decoding method that incorporates medical information, acquired from the training data, into the beam search of the diagnostic text generation process. We evaluate our method on two medical datasets using four Diagnostic Captioning systems that span from image-to-text systems based on convolutional encoders and recurrent network decoders to pre-trained Large Language Models (LLMs). The latter can also be used in few-shot and zero-shot learning scenarios. In most cases, when using our proposed decoding mechanism, the performance of beam search was enhanced with respect to all four evaluation metrics.

1 Introduction

Diagnostic Captioning (DC) systems receive the medical examinations of a patient, such as X-Rays or Magnetic Resonance Images (MRIs), which they analyse in order to draft a diagnostic report describing the medical condition of that patient (Ting et al., 2023). Such systems can function as supportive tools for doctors and clinical staff, assisting them in their daily workload. Although relatively unclear, the research goals of DC were summarized by Pavlopoulos et al. (2021) as (i) increased overall throughput of medical departments since improving a partially correct draft report may be faster than writing it from scratch, (ii) reduced diagnostic errors by providing suggestions for the clinical findings of the input image, and (iii) decreased cost of medical imaging examinations. Despite the rapid advancements in deep learning methods, the draft diagnostic reports generated by DC systems still exhibit shortcomings, such as hallucinations or the lack of the accurate description of the conditions depicted in the image (Xu et al., 2023). Although the former can be partly addressed with larger language models , the accuracy of a generated diagnostic text strongly depends on whether key medical terms related to the patient's examination were considered during text generation (Huang et al., 2019; Wang et al., 2020). Such key terms can be retrieved from the patient's medical history or generated by a medical image classifier and are widely referred to as medical "concepts" or "tags" (Pavlopoulos et al., 2019). Assigning concepts to an image is to some extent similar to content selection, which boils down to deciding which tokens to generate at a conceptual or logical level and was the first stage in older symbolic-based text generation systems (Reiter and Dale, 2000).

To this end, we propose "Distance from Median Maximum Concept Similarity" (DMMCS), a novel guided decoding framework that aims to integrate key biomedical terms associated with the corresponding medical examinations into the diagnostic text generation process. This is achieved by imposing a custom case-specific penalty at each decoding step. The penalty is designed to prioritize the generation of words that are contextually similar with the relevant medical tags. In this way, we manage to integrate information from key biomedical terms regarding the patients' condition into the generated diagnostic reports. Thus, the semantic affinity, i.e., the alignment or resemblance in meaning, between the generated and the ground truth report can be

enhanced. DMMCS involves calculating a series of statistical distributions that model the relationship between each concept and the tokens of the diagnostic captions with which it is associated in the training data. To the best of the authors' knowledge, this is the first guided decoding method specifically developed for DC systems, as well as the first using information from image concepts to guide the generation of an image caption.

DMMCS is applicable to any DC system, as it can be integrated in the decoding process of any encoder-decoder model responsible for the text generation. For experimental purposes, we trained four DC systems, spanning from a conventional Show&Tell method (Vinyals et al., 2015), to Transformer-based architectures (Vaswani et al., 2017), and state-of-the-art prompt-based systems (Dai et al., 2023; Alayrac et al., 2022). Furthermore, we investigated the impact of the DMMCS framework on few-shot captioning scenarios. We evaluated the performance of all models across two datasets from the medical domain, namely the ImageCLEFmedical 2023 dataset (Rückert et al., 2023) and MIMIC-CXR (Johnson et al., 2019). We used four diverse evaluation metrics, namely BLEU, ROUGE, BERTscore and BLEURT, comparing the results with and without the application of the DMMCS-based decoding. This allowed us to assess and demonstrate the effectiveness of DMMCS-based decoding in enhancing the models' performance in DC.

Overall, our main contributions are: (i) We introduced a guided decoding method designed specifically for DC systems. It leverages a set of medical tags associated with the respective image, aiming to enhance the accuracy and relevance of the generated diagnostic captions with regard to the ground truth diagnosis. (ii) We trained four DC systems covering a wide range of learning scenarios, including fine-tuning and few-shot learning. (iii) We demonstrated that applying the DMMCS-based decoding method significantly enhances performance across all four models in the vast majority of cases.

2 Related Work

A considerable amount of research has been dedicated to Controllable Text Generation and guided decoding strategies (Prabhumoye et al., 2020). Standard decoding methods, such as vanilla beam or greedy search, provide minimal control over the model's output (Zhou et al., 2023). Therefore, decoding techniques that partially guide the model's choices to adhere to task-specific requirements have been proposed.

The common objectives in guided decoding methods can be categorized based on three control conditions: semantic, structural and lexical (Zhang et al., 2023). Semantic constraints guide the model to generate text conforming to specific attributes such as tense or sentiment (Yang et al., 2018; Gu et al., 2022). For instance, Ghazvininejad et al. (2017) introduced *Hafez*. It was initially designed for generating poetry-styled texts, but is adaptable to meet any specified content-based decoding constraint. This is achieved by integrating a constraint-specific score in the next word selection process at each decoding step. The constraint-specific scores are calculated by a feature function. Such functions can, for example, encourage or discourage specific word choices, prevent repetitions, or guide the model to prefer longer words. The constraints may also be based on supervised learning (Holtzman et al., 2018) or reinforcement learning (Li et al., 2017) rather than just heuristics (Baheti et al., 2018). Structural constraints direct the model's output to adhere to a specific syntax structure (Yang et al., 2022). (Chen et al., 2019; Kumar et al., 2020). Finally, lexical-based constraints guide the model's choices to incorporate a set of specified keywords. He (2021) introduced CBART (constrained BART), which shifts a part of the generation process from the decoder to the encoder. The encoder guides the generation towards some specified must-include tokens. This is achieved by adding a dense layer over BART's encoder (Lewis et al., 2020) that generates a sequence of labels. The latter guides the decoder on which actions should be taken. The model undergoes a refinement process, regenerating multiple outputs until all constraints are fulfilled.

A recent novel method for guided decoding, is *Contrastive Search*, introduced by Su et al. (2022). It attempts to tackle the problem of text degeneration, where models produce unnatural and repetitive text. The authors proposed imposing a degeneration penalty at each decoding step to guide the model's output towards more natural and less repetitive text sequences. The degeneration penalty for a candidate token u at decoding step t is defined as follows:

$$D_u = \max_{1 \leq j \leq t-1} sim\left(h(u), h(x_j)\right), \qquad (1)$$

where x_j represents the preceding tokens of the incomplete generated sequence x, while the function $h(\cdot)$ calculates the word embeddings. Moreover, $sim(\cdot,\cdot)$ denotes the cosine similarity function, i.e., $sim(x,y) = \frac{x \cdot y}{\|x\|\|y\|}$. The degeneration penalty is calculated as the maximum cosine similarity between the word embeddings of the candidate token u and the preceding sequence $x_{<t}$ (denoted as $h(u)$ and $h(x_t)$ respectively). It negatively affects the overall score of each candidate token u at the specific decoding step. Thus, larger penalty values will guide the decoding algorithm to address the repetition and improve word variation, by preferring less repetitive words in the output.

3 The DMMCS Method

Our proposed method introduces a concept-guided decoding strategy for DC systems. It aims to guide the model in selecting words that are semantically similar to the tags (medical terms) associated with the respective images. This is achieved by imposing a custom penalty at each decoding step. The tags outline the key aspects of the patient's medical condition. Hence, integrating information from them into the diagnostic caption can lead to enhanced accuracy of the draft report. For instance, if a radiology image is associated with the tag "Atelectasis", but the generated caption makes no implicit or explicit reference to the aforementioned medical condition, then it is potentially inaccurate.

As a first step, we calculated the FastText word embeddings (Bojanowski et al., 2017) of both the tags and each token in the captions. Tags consisting of multiple tokens were represented by the centroid of the tokens' word embeddings. We then investigated the relationship between the training captions and the gold tags with which they were associated. Figure 1 presents a heatmap which visualizes the relationship between the tokens of a caption s (on the x-axis) and its corresponding set of tags T (on the y-axis). Each heatmap block represents the cosine similarity between the (centroid) embedding of the corresponding tag t and the respective caption token s_j. Darker colors correspond to larger similarity values. Hence, for instance, the lower right block denotes that the 11-th and 12-th tokens s_{11}, s_{12} share a similar meaning with the corresponding tag t_3. Thus, t_3 is prominently represented in the caption s, while the rest of the tags are not so strongly featured. We define the relationship between a tag t and a caption s as

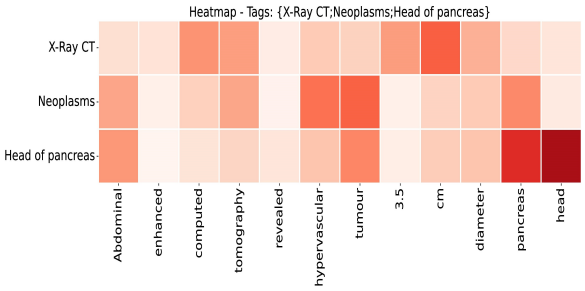

Figure 1: Heatmap visualizing the cosine similarities between the tokens of a ground truth caption (x-axis) and its associated biomedical tags (y-axis).

the *maximum cosine similarity (MCS)* between the (centroid) word embedding of t and the word embedding of each token in s. Mathematically, MCS is calculated as:

$$MCS(t,s) = \max_{1 \le j \le |s|} sim(h(t), h(s_j)), \quad (2)$$

where $sim(\cdot,\cdot)$ and $h(\cdot)$ denote the cosine similarity and the embeddings generation function respectively, as defined in Eq. 1. A high MCS value between a tag and a caption indicates a significant presence of the tag's meaning within this caption.

Next, we decided to investigate the relationship between a tag and all the captions with which it is associated in the training data. For a particular tag t, given its associated captions S, we calculate its MCS distribution R, as the following set:

$$R(t,S) = \{MCS(t,s) | s \in S\}, \quad (3)$$

where $MCS(t,s)$ is defined in Eq. 2.

Figure 2 illustrates the MCS distribution R of the tag "Angiogram", present in the ImageCLEF 2023 dataset (Rückert et al., 2023). The blue box represents the interquartile range (IQR) of the distribution, while the coral line denotes the median value based on the training data.

We repeated the calculation of the *MCS* distribution R for every tag t in the training dataset. Figure 3 presents the interquartile range (IQR) per tag t, for the ImageCLEF dataset, reflecting how strongly t is represented among the captions it is associated with. The median maximum cosine similarity (*MMCS*) per IQR forms a coral line.
The *MMCS* value for a tag t, given the set of its associated training captions S, is calculated as:

$$MMCS(t,S) = median(R(t,S)), \quad (4)$$

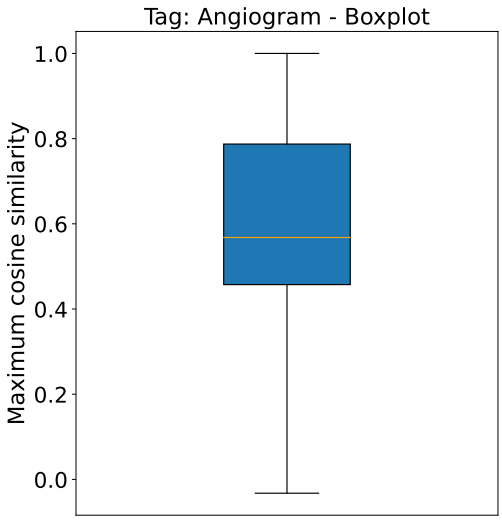

Figure 2: A boxplot representing the MCS distribution R for the tag "Angiogram".

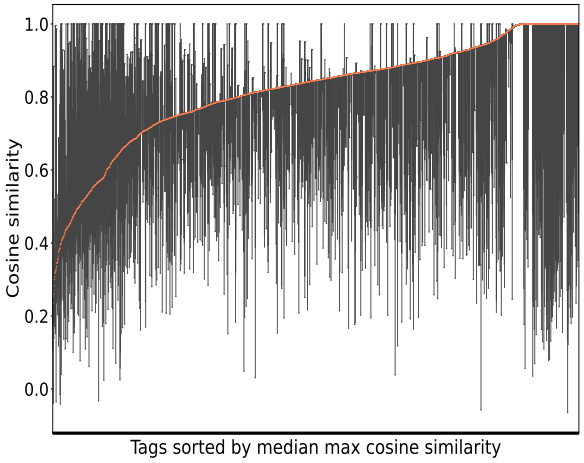

Figure 3: IQR plot of the *MCS* distribution per tag, sorted based on the *MMCS* value (the median, in coral).

where R denotes the MCS distribution of the tag, as defined in Eq. 3. It quantifies the average lexical representation of the tag in its associated captions. The plots are sorted in ascending order based on their median value (and thus the respective tag's *MMCS* value), in order to highlight the observation that the tags in the ImageCLEFmedical 2023 dataset are not consistently strongly expressed in the training ground truth captions (also, see §4). For instance, we observe a variation in *MMCS* values across the plot, ranging from as low as 0.3 for tags on the left end, to values close to 1 for tags on the right end. The former are overall implicitly represented in the diagnostic captions with which they are associated, while the latter are consistently explicitly mentioned.

Considering these findings, we propose a new penalty that aims to improve the generation of diagnostic captions by effectively integrating information provided by the associated biomedical concepts. By applying this penalty at each decoding step, the decoder is prompted to choose words that more or less explicitly convey the associated tags. The level of explicitness is determined by the *MMCS* value of each tag, as computed in the training captions. Hence, tags with larger *MMCS* values should be consistently mentioned, while tags with lower *MMCS* values should be expressed less explicitly.

During inference, when an image is associated with a single tag, we initially calculate the *MCS* between the tag and the tokens of the caption that have already been generated (see Eq. 2). The penalty is then defined as the squared difference between the computed *MCS* value, which represents how strongly the term is expressed in the current caption and the tag's *MMCS* value, which corresponds to how strongly the tag is expressed (median value) in the gold training captions associated with the tag. In case an image is associated with multiple tags, then an individual penalty is calculated for each one, as described before and the total penalty is equal to their sum divided by the number of associated tags. Formally, given a (possible still incomplete) caption s and a set of associated tags T, the penalty is calculated as:

$$DMMCS_p(T, s) = \frac{1}{|T|} \cdot \sum_{t \in T} (MCS(t, s) - MMCS(t))^2 \quad (5)$$

In essence, the penalty favors a candidate caption s whose *MCS* score (across the tokens of s and each tag t) falls aggregately closer to the *MMCS* score of each tag t. The former reflects the maximum lexical representation of the tag in the words of the candidate caption, while the latter denotes the average lexical representation of the tag in the training captions it is associated with in the training set. Intuitively, the objective is to generate a caption such that each tag is expressed as strongly (close to the median strength) as in the training captions of the tag. Additionally, there are two scenarios regarding the source of the biomedical tags;

they are either provided (ground truth) or generated by a biomedical image classifier.

At each decoding step, the sampling algorithm considers the n best candidate captions, according to the following score:

$$DMMCS(s) = \alpha \cdot (1 - HD) \cdot DMMCS_p(T, s) + (1 - \alpha) \cdot HD \cdot (1 - D_{score}), \quad (6)$$

where s is is a (possibly still incomplete) candidate generated caption considered by the beam search decoder and T a set of tags the input image (to be captioned) is associated with. The rest of the terms are explained below.

***D_{score}*:** This is the score that the decoder of the model assigns to each candidate sequence, that is the sum of the (negative log) probabilities of the decoder (Eq. 7). The decoder can be any type of generative model conditioned on the input image, e.g., a Recurrent Neural Network (RNN) or a Transformer-based architecture (Vaswani et al., 2017). We employed experiments with both types of architectures in order to test the effectiveness of the DMMCS-based decoding. The D_{score} is also normalized in $[0, 1]$, using min-max scaling, in order to align with the $DMMCS_p$ score's range. As the goal is to minimize the overall score, the formula in Eq. 6 incorporates the minimization of $1 - D_{score}$.

$$D_{score} = -\sum_{t=1}^{t} \log P(s_t | s_{<t}) \quad (7)$$

Weighting factor α: The parameter α acts as a regulatory factor aiming to statically control the effect of the two addends in the overall score. The larger the value of α, the more significant the influence of the $DMMCS_p$ on the score, and vice versa. Therefore, when $\alpha = 0$, normal beam search is applied, while when $\alpha = 1$, we solely trust the DMMCS mechanism. This parameter is fixed and determined after experimenting with several values on the validation set. It is also used in order to scale the value of the two addends, in order to ensure that both will substantially contribute to the overall score.

Histogram Divergence (HD): This is a dynamically computed parameter that serves as a weighting factor of the two components; the $DMMCS_p$ and the D_{score}. At each decoding step, we consider a set of generated candidate sequences G. A histogram of the MCS distribution $R(t, G)$ for each tag t is calculated. It represents the maximum lexical representation of the tag t in each candidate sequence. Therefore for a given list of ground truth tags T, $|T|$ histograms would be generated. In addition, a similar histogram has already been calculated for each tag on its associated ground truth captions with respect to the training data. The rationale is that beams with *MCS* distributions (one for each tag) similar to those computed on the training data should assign a larger weight to the $DMMCS_p$ factor compared to the D_{score} during that decoding step. In contrast, if the distributions do not match, the typical D_{score} should be trusted more.

Given a single tag c, the divergence of its two corresponding histograms (computed on the generated and ground truth captions respectively) is calculated by performing a Kolmogorov-Smirnov Goodness of Fit Test (KS-test) (Massey, 1951), which yields two values: the p-value and the ks-statistic. The p-value, as in any other goodness of fit test, can be examined to either accept or reject a null hypothesis but it does not precisely quantify the discrepancy between the two provided distributions. Unlike p-value, the ks-statistic measures the distance between two given distributions, serving as an appropriate metric in the specified context as it takes values in $[0, 1]$. A value of 0 suggests that the two samples are drawn from the same distribution, while a value of 1 indicates the opposite. The ks-statistic is calculated as the maximum absolute vertical distance between the Empirical Cumulative Distribution Functions (ECDF) of the two distributions. Formally, the ks-statistic for a single tag t at a random decoding step can be calculated as:

$$ks(t) = \max |F(R(t, S)) - F(R(t, G))|, \quad (8)$$

where S denotes the training captions associated with t and G represents the generated captions up to this decoding step. Moreover, $F(\cdot)$ calculates the ECDF of the given input, while R is the *MCS* distribution as defined in Eq. 3. Overall, the HD for a set of tags T is calculated as:

$$HD = \frac{1}{|T|} \cdot \sum_{t \in T} ks(t). \quad (9)$$

In summary, the proposed method operates as follows. At each decoding step, the $DMMCS(s)$

score (Eq. 6) is calculated for each candidate sequence s generated through the beam search. The score requires the computation of HD (Eq. 8) and $DMMCS_p$ (Eq. 5). The former signifies the amount of trust that is put into the latter, which measures how strongly the examined image's tags are represented in the generated caption. The n sequences with the lowest scores are selected and considered throughout the next decoding step.

4 Experiments

In this section, we discuss the key experiments conducted to assess and validate the effectiveness of the DMMCS penalty for DC. We implemented four DC models, spanning diverse architectures and learning techniques, including long-established image captioning models and advanced state-of-the-art approaches. Moreover, we evaluated the proposed algorithm using two distinct datasets; the dataset of the ImageCLEFmedical 2023 campaign (Rückert et al., 2023) and a subset of the MIMIC-CXR dataset (Johnson et al., 2019). The performance of each model is evaluated using four metrics, briefly introduced in §4.3. We demonstrate that the integration of our proposed guided decoding method in caption generation leads to significant improvement in the model's performance in the vast majority of cases.

4.1 Datasets

ImageCLEFmedical 2023 Dataset: The ImageCLEFmedical 2023 dataset (Rückert et al., 2023) is a subset of the Radiology Objects in Context (ROCO) dataset (Pelka et al., 2018). It consists of 71,355 images, accompanied by their respective biomedical tags, as well as a brief diagnostic caption. It contains 2,125 distinct biomedical tags, and the images belong to multiple medical modalities such as X-Ray, CT, and MRI. The dataset, as part of the annual ImageCLEF campaign (Ionescu et al., 2023), does not dispose ground truth captions for the test set. Consequently, we assessed the model's performance on a held-out subset.

MIMIC-CXR Dataset: MIMIC-CXR (Johnson et al., 2019) originally consists of 377,110 deidentified chest X-ray examinations. Each of them is accompanied by a set of biomedical tags and a brief diagnostic caption. There are 14 distinct biomedical tags in total, generated by from the CheXpert labeler system (Irvin et al., 2019). We only considered examinations containing a single radiology image of the patient, because the DC systems we used could only receive a single image as input. Thus, the subset consisted of 32,174 images along with their associated biomedical tags and captions. We randomly split them into training, validation and test sets.

4.2 DC Models

The following four DC Models were used for the evaluation of the proposed algorithm.

Show & Tell: This model is known as Show & Tell architecture (Vinyals et al., 2015) or CNN-RNN, as it consists of a Convolutional Neural Network (CNN) image encoder followed by an RNN text generation module. The CNN encoder is in charge of extracting image features that are then fed to the RNN decoder, which is responsible for generating the diagnostic caption based on these features.

ViT-GPT2: This system is also based on the Encoder-Decoder approach, except that now both components are based on transformers (Vaswani et al., 2017). Vision Transformer (ViT) (Dosovitskiy et al., 2021) is employed as the image encoder, while GPT2 (Radford et al., 2019) is responsible for caption generation. Both models were loaded from a checkpoint on HuggingFace Hub[1] for a joint ViT-GPT2 encoder-decoder pipeline, further pre-trained on generic image-caption pairs and then fine-tuned on the two biomedical datasets.

InstructBLIP: This is a vision-language instruction-tuning model introduced by (Dai et al., 2023). Such models are designed to learn and swiftly adapt to new tasks based on specific instructions provided during the training process. Hence, their performance is strongly dependent on the provided instruction.

Flamingo: Flamingo (Alayrac et al., 2022) is a few-shot in-context image captioning system. Flamingo can generate a diagnostic caption based on a few demonstrative examples of image-caption pairs provided as a multi-modal prompt. Due to the fact that checkpoints of the original Flamingo architecture are not publicly available, we employed OpenFlamingo, an open-source implementation that obtains around 80-89% of the corresponding Deep-

[1] https://huggingface.co/nlpconnect/vit-gpt2-image-captioning

Mind's Flamingo model's performance (Awadalla et al., 2023).

4.3 Evaluation Metrics

We provide a brief introduction of the four evaluation metrics employed in order to assess the proposed algorithm: BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), BERTscore (Zhang et al., 2020), and BLEURT (Sellam et al., 2020). The first two are word-based, focusing on the lexical overlap between the machine-generated sequence and the ground truth sequence. On the other hand, BERTscore and BLEURT emphasize on capturing the contextual similarity between two texts. This involves measuring the degree of similarity in meaning and context, considering the overall understanding and semantics of the provided sentences.

4.4 Results

In this section, we present a quantitative and qualitative analysis on the results of our experiments with the proposed decoding algorithm. Moreover, we assess each model's performance with and without the integration of the DMMCS-based decoding. We repeat every experiment for three random non-intersecting samples of the test set and report the average score for each metric. Moreover, we use the ground truth tags as the set of keywords for the decoding algorithm.

Tuning of the weighting factor α: The effect of the proposed penalty ($DMMCS_p$) and the decoder score for each candidate sequence (D_{score}) is regulated by the weighting factor α, as shown in Eq. 6. Thus, we performed a tuning process in order to decide the optimal value of α for each model. Specifically, we evaluated every model's performance on a held-out set for multiple values of α, with an extensive search ranging from 0.05 to 0.95. Figure 4 presents the performance of InstructBLIP (Dai et al., 2023) on the ImageCLEFmedical 2023 dataset when the DMMCS-based decoding method is applied. It shows four plots, one for each evaluation metric, which illustrate how the corresponding scores fluctuate as the α value steadily increases. The red dotted line acts as a baseline and denotes the performance of the model when standard beam search is applied. We observe that there is no specific range of α values for which the model consistently outperforms the baseline (standard beam search) on all evaluation metrics. Specifically, ROUGE and BERTscore follow a specific pattern, where values of α up to 0.4 prove beneficial for the model's performance. However, BLEU and BLEURT do not adhere to this pattern, as larger α values ($\alpha > 0.4$) result in improved performance, while smaller values lead to lower scores as it is illustrated below. Additional plots regarding the selection of the optimal weighting factor α for each model, along with the method's performance improvements in various scenarios, can be found in Appendix A.

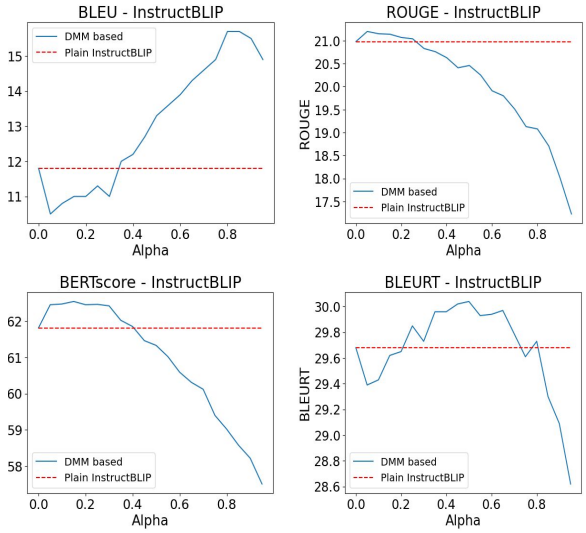

Figure 4: The InstructBLIP model's performance for various α values on ImageCLEF medical 2023 dataset (Ionescu et al., 2023).

Quantitative results: Table 1 provides an overview of the performance of each model across the four evaluation metrics averaged on the three test subsets. For each evaluation metric, two scores are provided that correspond to beam search (denoted as *BS*) and DMMCS-based decoding respectively. The scores for DMMCS are obtained with the best α value for each metric based on the search using the validation set. We observe that for each model, for each dataset, *DMMCS* outperforms beam search consistently in most of the metrics. When benefits were not recorded, the margin was small (-0.2 or less). By contrast, the improvements were of a larger margin (up to 4 points in the third row of sub-table A). In addition, we conducted experiments using the DMMCS-based decoding with tags predicted by a medical image classifier instead of the gold tags. Our results, which are discussed in Appendix B, show that model performance is again improved compared to the standard beam search decoder.

A. ImageCLEFmedical 2023 Dataset								
	BLEU		ROUGE		BERTscore		BLEURT	
	BS	*DMMCS*	*BS*	*DMMCS*	*BS*	*DMMCS*	*BS*	*DMMCS*
Show & Tell	20.61	**21.48**	22.00	**23.01**	60.07	**60.58**	29.99	**30.53**
ViT-GPT2	15.34	**16.72**	16.72	**17.05**	59.38	**59.97**	26.50	**27.69**
InstructBLIP	11.81	**15.74**	20.98	**21.20**	61.81	**62.54**	29.68	**30.04**
Flamingo	15.34	**15.93**	15.98	**16.21**	**55.36**	55.21	28.49	**28.89**
B. MIMIC-CXR Dataset								
	BLEU		ROUGE		BERTscore		BLEURT	
	BS	*DMMCS*	*BS*	*DMMCS*	*BS*	*DMMCS*	*BS*	*DMMCS*
Show & Tell	11.77	**14.13**	**17.18**	17.10	58.49	**58.82**	29.08	**29.49**
ViT-GPT2	13.55	**14.78**	21.37	**21.74**	**56.77**	56.71	23.52	**24.37**
InstructBLIP	12.76	**13.32**	14.35	**16.27**	58.59	**59.12**	24.65	**25.56**
Flamingo	12.78	**13.26**	13.68	**13.86**	**51.45**	51.28	29.14	**29.81**

Table 1: The performance of each model on both datasets evaluated across four metrics. *BS* and *DMMCS* denote beam search and DMMCS-based decoding respectively. For the DMMCS-based decoding, gold tags provided in the training set are used.

Qualitative results: Additionally, we conducted a qualitative analysis of the diagnostic captions produced by the DC models with and without applying our proposed guided decoding method. We present a sample of the findings in Figure 5. Without DMMCS, the beam search decoder generated a partially inaccurate caption, referring to a "parasternal long axis view" instead of an "echocardiogram". While related, these terms are not precisely the same. Given the importance of precise medical terminology in diagnostic captions, addressing such inaccuracies is crucial. The DMMCS-enhanced caption rectified the examination type to an "echocardiography parasternal long axis view" aligning more closely with the reference caption. Moreover, it correctly identified the diagnosed medical condition as "pericardial effusion" without introducing extra inaccurate information compared to beam search decoding. However, it is crucial to mention that the proposed method may not address all hallucinations, or more generally, inaccuracies of the model. Nevertheless, the draft reports generated by our method are more accurate according to all evaluation metrics, while these draft captions will be checked and improved by medical experts.

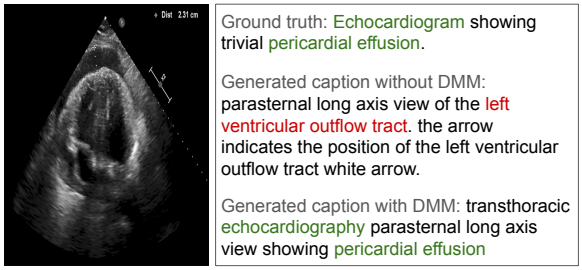

Figure 5: A sample of the findings of an exploratory qualitative analysis on captions generated by a ViT-GPT2 instance using standard beam search and DMMCS-enhanced decoding respectively.

5 Conclusion

This work introduced DMMCS, a guided decoding method that enhances the accuracy of (automatically-generated) diagnostic captions by integrating medical tags associated with a radiology image in the generation process. We assessed the effectiveness of the proposed method by applying it to four systems with diverse architectures and on two datasets, viz. ImageCLEFmedical 2023 and a subset of MIMIC-CXR. Subsequently, we compared the accuracy of the generated captions with those obtained using a typical beam search decoding approach, employing four widely used evaluation metrics. Our results demonstrate that the proposed DMMCS mechanism consistently outperforms typical beam search decoding across almost all models and datasets for most of the metrics.

In future work we plan to experiment with specific medical domains and focus on a broader range of tasks to investigate the benefits of our method in a wider context. Furthermore, we will explore our method's capabilities in generic image captioning (Lin et al., 2015), as well as other text generation tasks, including question-answering, text summarization, and machine translation. A final direction for future work concerns the use of contextual representations, so that the tags (and captions) will become context-aware, to investigate whether different occurrences of the same tag in a caption carry different information.

6 Limitations

- The generalisability of the proposed method is limited by the constrained number of datasets used in this study. Although we used all the publicly available medical datasets, to the best of our knowledge, the results are limited to the specific conditions and regions (and language, i.e, English) these datasets concern. This limitation could be addressed by collaborating with medical institutions, under the license of respective review boards, which is a direction we plan to consider in the future, building on the work presented in this article.

- The success of the proposed method depends on the quality of the medical tags. This means that noisy input (tags) could affect the performance. To investigate thishypothesis, besides the gold tags, we also used tags retrieved by an image classifier for the ImageCLEF medical 2023 dataset. As it is illustrated at the top rows of Table 1 (§4.4), the increase in performance with *DMMCS*-based decoding when using the predicted tags slightly drops.

- Clinical correctness of automatically-generated diagnostic reports is critical. Although our *DMMCS*-based generated reports were more accurate than the baseline, we note that their clinical correctness has not been assessed. That is mainly due to the absence of appropriate evaluation measures, which is a direction that future research should focus on.

6.1 Ethical considerations

Our proposed *DMMCS*-based decoding method is related to the 3rd and the 10th goal of United Nations Sustainability Goals, since assisting doctors and clinicians by providing them more accurate draft diagnostic reports promotes ensuring good health and well-being, as well as reduced inequalities. However, emphasis on biomedical data privacy has long been a sensitive issue because of crossing ethical, legal, and technical boundaries, thus apprehension of clinical information privacy needs to be taken into serious consideration when patients' data are used for model training.

Moreover, when gold tags are not provided and need to be retrieved from a patient's medical history, developed countries, where medical information is typically stored in large information systems, have a clear advantage over undeveloped regions or countries under development. In these situations, a medical image classifier can be employed to predict the medical concepts depicted in an image, as it is demonstrated in Appendix B. We believe that the positive societal impact associated to using our method exceeds potential problems and limitations of DC systems in general.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736, New Orleans, LA, USA.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. *arXiv preprint arXiv:2308.01390*.

Ashutosh Baheti, Alan Ritter, Jiwei Li, and Bill Dolan. 2018. Generating More Interesting Responses in Neural Conversation Models with Distributional Constraints. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3970–3980, Brussels, Belgium. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019. Controllable Paraphrase Generation with a Syntactic Exemplar. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5972–5984, Florence, Italy. Association for Computational Linguistics.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference*

on Computer Vision and Pattern Recognition, pages 248–255, Miami, FL, USA.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, 2021*.

Marjan Ghazvininejad, Xing Shi, Jay Priyadarshi, and Kevin Knight. 2017. Hafez: an Interactive Poetry Generation System. In *Proceedings of ACL 2017, System Demonstrations*, pages 43–48, Vancouver, BC, Canada. Association for Computational Linguistics.

Yuxuan Gu, Xiaocheng Feng, Sicheng Ma, Lingyuan Zhang, Heng Gong, and Bing Qin. 2022. A Distributional Lens for Multi-Aspect Controllable Text Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1023–1043, Abu Dhabi, United Arab Emirates.

Xingwei He. 2021. Parallel Refinements for Lexically Constrained Text Generation with BART. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8653–8666, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. 2018. Learning to Write with Cooperative Discriminators. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1638–1649, Melbourne, Australia. Association for Computational Linguistics.

Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. pages 4700–4708.

Xin Huang, Fengqi Yan, Wei Xu, and Maozhen Li. 2019. Multi-Attention and Incorporating Background Information Model for Chest X-Ray Image Report Generation. *IEEE Access*, 7:154808–154817.

Bogdan Ionescu, Henning Muller, Ana-Maria Dragulinescu, Wen-wai Yim, Asma Ben Abacha, Neal Snider, Griffin Adams, Meliha Yetisgen, Johannes Ruckert, Alba Garcia Seco de Herrera, Christoph M. Friedrich, Louise Bloch, Raphael Brungel, Ahmad Idrissi-Yaghir, Henning Schafer, Steven A. Hicks, Michael A. Riegler, Vajira Thambawita, Andrea Storås, Pål Halvorsen, Nikolaos Papachrysos, Johanna Schöler, Debesh Jha, Alexandra-Georgiana Andrei, Ahmedkhan Radzhabov, Ioan Coman, Vassili Kovalev, Alexandru Stan, George Ioannidis, Hugo Manguinhas, Liviu-Daniel Stefan, Mihai Gabriel Constantin, Mihai Dogariu, Jerome Deshayes, and Adrian Popescu. 2023. Overview of ImageCLEF 2023: Multimedia retrieval in medical, socialmedia and recommender systems applications. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Proceedings of the 14th International Conference of the CLEF Association (CLEF 2023), Thessaloniki, Greece. Springer Lecture Notes in Computer Science LNCS.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Alistair Johnson, Tom Pollard, Seth Berkowitz, Nathaniel Greenbaum, Matthew Lungren, Chih-ying Deng, Roger Mark, and Steven Horng. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6:317.

Ashutosh Kumar, Kabir Ahuja, Raghuram Vadapalli, and Partha Talukdar. 2020. Syntax-Guided Controlled Generation of Paraphrases. *Transactions of the Association for Computational Linguistics*, 8:329–345.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2017. Learning to Decode for Future Success.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft COCO: Common Objects in Context.

Frank J. Massey. 1951. The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, 46(253):68–78.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

John Pavlopoulos, Vasiliki Kougia, and Ion Androutsopoulos. 2019. A Survey on Biomedical Image Captioning. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 26–36, Minneapolis, Minnesota. Association for Computational Linguistics.

John Pavlopoulos, Vasiliki Kougia, Ion Androutsopoulos, and Dimitris Papamichail. 2021. Diagnostic Captioning: a Survey. *Knowledge and Information Systems*, 64:1691 – 1722.

Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christopher Friedrich. 2018. Radiology Objects in COntext (ROCO): A Multimodal Image Dataset. In *CVII-STENT/LABELS@MICCAI*.

Shrimai Prabhumoye, Alan W Black, and Ruslan Salakhutdinov. 2020. Exploring Controllable Text Generation Techniques. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1–14, Barcelona, Spain (Online).

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Studies in Natural Language Processing. Cambridge University Press.

Johannes Rückert, Asma Ben Abacha, Alba Herrera, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, Henning Schäfer, Henning Müller, and Christoph Friedrich. 2023. Overview of ImageCLEFmedical 2023 – Caption Prediction and Concept Detection. In *CLEF2023 Working Notes*, CEUR Workshop Proceedings, Thessaloniki, Greece. CEUR-WS.org.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A Contrastive Framework for Neural Text Generation. In *Advances in Neural Information Processing Systems*, volume 35, pages 21548–21561, New Orleans, LA, USA.

Pang Ting, Peigao Li, and Lijie Zhao. 2023. A Survey on Automatic Generation of Medical Imaging Reports Based on Deep Learning. *BioMedical Engineering OnLine*, 22.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6000–6010, Long Beach, California, USA.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A Neural Image Caption Generator. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, Boston, MA, USA.

Fuyu Wang, Xiaodan Liang, Lin Xu, and Liang Lin. 2020. Unifying Relational Sentence Generation and Retrieval for Medical Image Report Composition. *IEEE Transactions on Cybernetics*, 52:5015–5025.

Liming Xu, Quan Tang, Jiancheng Lv, Bochuan Zheng, Xianhua Zeng, and Weisheng Li. 2023. Deep Image Captioning: A Review of Methods, Trends and Future Challenges. *Neurocomputing*, 546:126287.

Zhixian Yang, Pengxuan Xu, and Xiaojun Wan. 2022. Diversifying Neural Text Generation with Part-of-Speech Guided Softmax and Sampling. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6547–6563, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. 2018. Unsupervised Text Style Transfer using Language Models as Discriminators. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023. A Survey of Controllable Text Generation Using Transformer-Based Pre-Trained Language Models. *Association for Computing Machinery*, 56(3):1–37.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTscore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Wangchunshu Zhou, Yuchen Eleanor Jiang, Ethan Wilcox, Ryan Cotterell, and Mrinmaya Sachan. 2023. Controlled Text Generation with Natural Language Instructions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 42602–42613.

A Results for various α values

In this part, we provide additional insight regarding the influence of the weighting factor α to the DMMCS-based generation performance across different models and datasets, including a variety of few-shot scenarios. As mentioned in §4.4, we measured the performance on a held-out set for multiple values of α, extensively searching among values in $[0.05, 0.95]$. There are no particular subsets of α values for which DMMCS-based decoding consistently outperforms the baseline on all four evaluation metrics. Specifically, lower values of α are more beneficial for ROUGE and BERTscore,

while larger values of α result in higher BLEU and BLEURT scores.

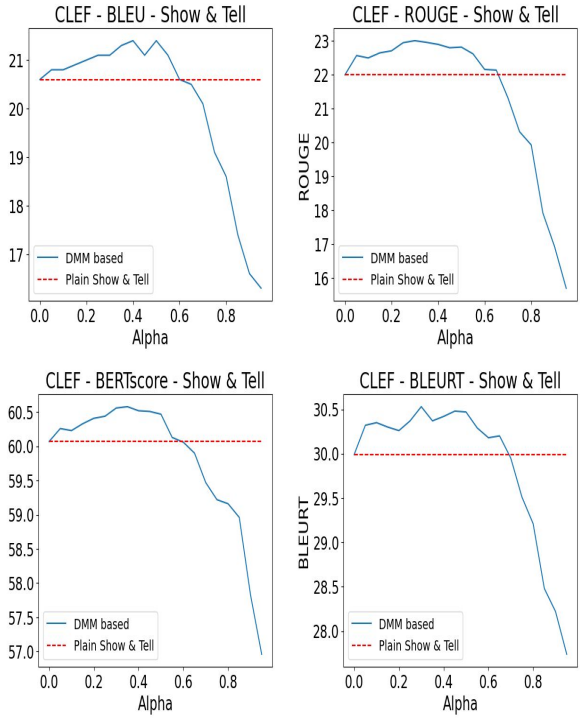

Figure 6: The Show and Tell model's performance for various α values on ImageCLEF medical 2023 dataset (Ionescu et al., 2023).

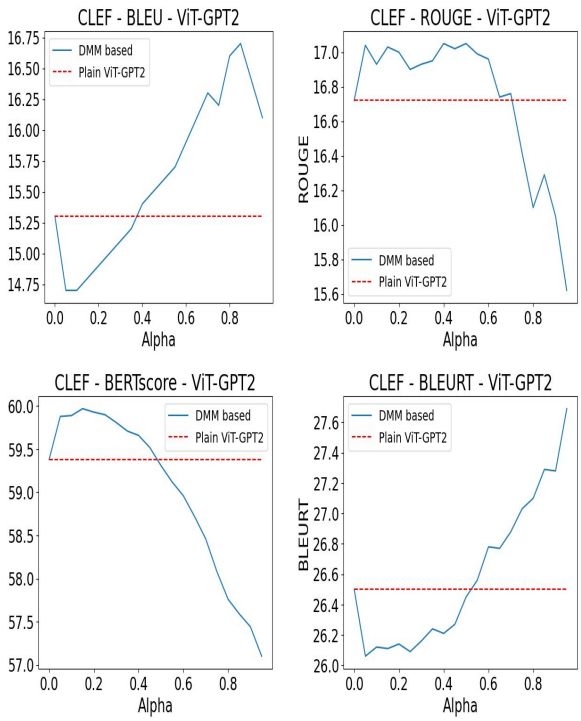

Figure 7: The ViT-GPT-2 model's performance for various α values on ImageCLEF medical 2023 dataset (Ionescu et al., 2023).

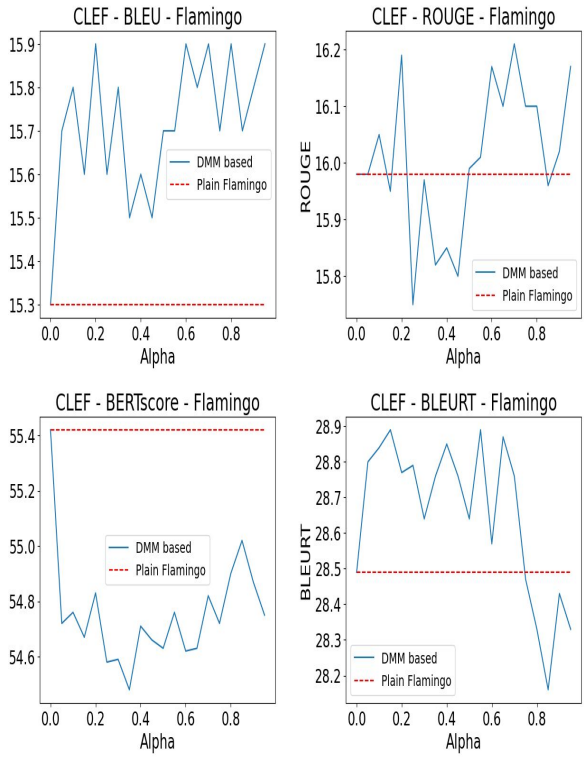

Figure 8: The OpenFlamingo model's performance for various α values on ImageCLEF medical 2023 dataset (Ionescu et al., 2023).

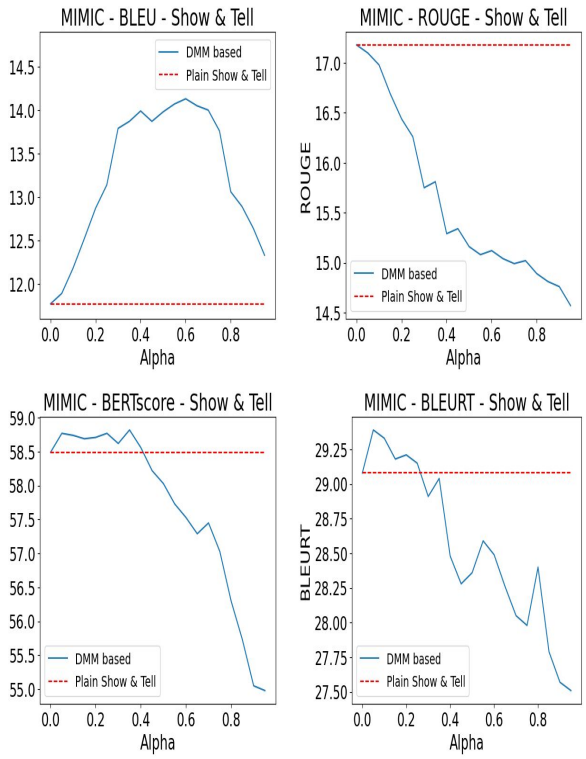

Figure 9: The Show and Tell model's performance for various α values on our selected subset of MIMIC dataset (Johnson et al., 2019).

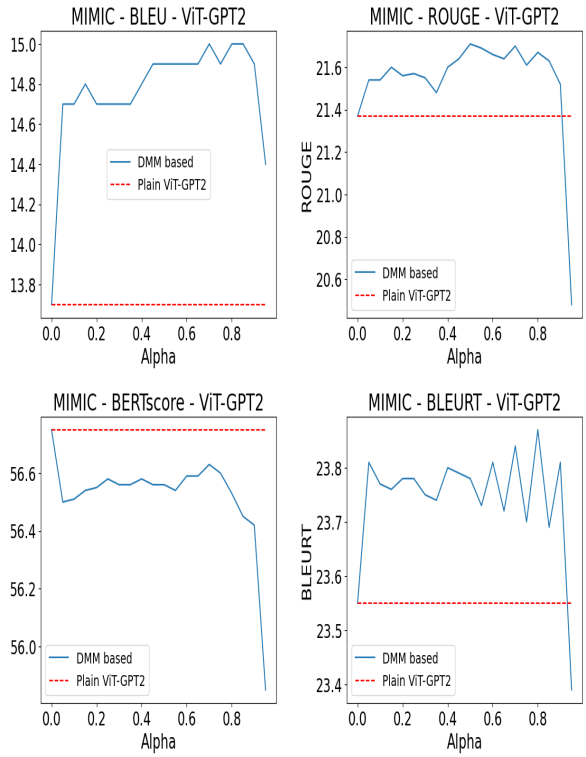

Figure 10: The ViT-GPT-2 model's performance for various α values on our selected subset of MIMIC dataset (Johnson et al., 2019).

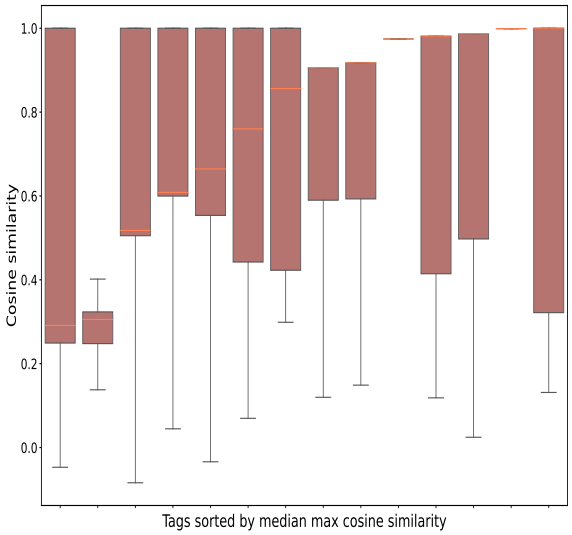

Figure 11: Multiple boxplots plotted side-by-side. Each plot visualizes the *MCS* distribution between a tag and its associated training captions in the MIMIC-CXR dataset. The boxplots are sorted based on their median value (coral line), which denotes the tag's median maximum cosine similarity (MMCS) value. Also see Figure 3 for the same plot on the ImageCLEF 2023 dataset.

B DMMCS based on tags predicted by a classification system

In this part, we provide the performance of the four models on the ImageCLEFmedical 2023 dataset (Rückert et al., 2023) obtained by employing an image classifier to acquire the tags used in DMMCS-based decoding. Specifically, we employed a DenseNet-121 instance (Huang et al., 2017), initially pre-trained on ImageNet (Deng et al., 2009), which we then fine-tuned on the biomedical domain using the ImageCLEFmedical 2023 dataset's image-tags pairs for the fine-tuning process. We observe that the performance of the models when the proposed algorithm is employed still outperforms the vanilla beam search ones. However, as expected, the increase in performance is less than when using the gold tags, as it is demonstrated in Table 2 and explained in §6.

C MIMIC-CXR MCS distributions

We also provide a figure with multiple interquartile plots for the employed MIMIC-CXR (Johnson et al., 2019) subset, in correspondance with Figure 3. Each plot represents the MCS distribution of each one of the 14 tags. The coral lines denote the median value of each interquartile plot (MMCS). We observe that, similarly to the ImageCLEFmedical 2023 dataset (see Figure 3), the MMCS value of the distinct tags vary significantly, ranging from 0.3 to 1.0.

ImageCLEFmedical 2023 Dataset - Predicted (not gold) Tags								
	BLEU		**ROUGE**		**BERTscore**		**BLEURT**	
	BS	*DMMCS*	*BS*	*DMMCS*	*BS*	*DMMCS*	*BS*	*DMMCS*
Show & Tell	20.61	**21.07**	22.00	**22.66**	60.07	**60.39**	29.99	**30.17**
ViT-GPT2	15.34	**16.11**	16.72	**16.84**	59.38	**59.44**	26.50	**27.01**
InstructBLIP	11.81	**13.98**	20.98	**21.01**	61.81	**62.31**	29.68	**29.76**
Flamingo	15.34	**15.47**	15.98	**16.03**	**55.36**	55.09	28.49	**28.63**

Table 2: The performance of each model on both datasets evaluated across the four metrics. *BS* and *DMMCS* denote beam search and DMMCS-based decoding respectively. For the DMMCS-based decoding, tags predicted by a medical image classifier are used.