A Data-Driven Guided Decoding Mechanism for Diagnostic Captioning

Panagiotis Kaliosis[1,2], **John Pavlopoulos**[1,2*], **Foivos Charalampakos**[1], **Georgios Moschovis**[1,2], **Ion Androutsopoulos**[1,2]

[1]Department of Informatics, Athens University of Economics and Business, Greece
[2]Archimedes/Athena RC, Greece
{pkaliosis, annis, phoebuschar, geomos, ion}@aueb.gr

Abstract

Diagnostic Captioning (DC) automatically generates a diagnostic text from one or more medical images (e.g., X-rays, MRIs) of a patient. Treated as a draft, the generated text may assist clinicians, by providing an initial estimation of the patient's condition, speeding up and helping safeguard the diagnostic process. The accuracy of a diagnostic text, however, strongly depends on how well the key medical conditions depicted in the images are expressed. We propose a new *data-driven* guided decoding method that incorporates medical information, in the form of existing tags capturing key conditions of the image(s), into the beam search of the diagnostic text generation process. We evaluate the proposed method on two medical datasets using four DC systems that range from generic image-to-text systems with CNN encoders and RNN decoders to pretrained Large Language Models. The latter can also be used in few- and zero-shot learning scenarios. In most cases, the proposed mechanism improves performance with respect to all evaluation measures. We provide an open-source implementation of the proposed method at https://github.com/nlpaueb/dmmcs.

1 Introduction

Diagnostic Captioning (DC) systems receive one or more medical images of a patient, such as X-Rays or Magnetic Resonance Images (MRIs), which they analyse to draft a diagnostic report (Ting et al., 2023). Such systems can function as supportive tools for doctors and clinical staff, assisting them in their daily workload. Possible benefits, as summarized by Pavlopoulos et al. (2021), include (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 images, which might otherwise be missed, and (iii) decreased cost of medical imaging examinations. Despite the rapid advancements in deep learning methods, draft diagnostic reports generated by DC systems still exhibit shortcomings, such as hallucinations or lack of accurate descriptions of medical findings (Xu et al., 2023). The medical accuracy of a generated diagnostic text strongly depends on whether key medical conditions depicted in the images are considered during text generation (Huang et al., 2019; Wang et al., 2020). Such key conditions can be captured by tags, reflecting medical concepts to be mentioned in the generated text. Tags of this kind can be obtained by medical image taggers (Rajpurkar et al., 2017; Lu et al., 2020) and are also present (as gold tags) in diagnostic captioning datasets. Assigning tags to an image is to some extent similar to content selection, i.e., deciding which concepts to express, which was the first stage in symbolic-based text generation systems (Reiter and Dale, 2000). More background information on DC is provided in Appendix A, and in the DC survey by Pavlopoulos et al. (2021).

In this work, we propose *Distance from Median Maximum Concept Similarity* (DMMCS), a novel data-driven guided decoding method that aims to integrate information from medical image tags into the diagnostic text generation process. This is achieved by imposing a new penalty at each decoding step. The penalty is designed to prioritize the generation of words that are semantically similar to the medical tags of the input images, also taking into account how often each tag is explicitly or implicitly expressed in gold captions. DMMCS involves calculating a series of statistical distributions that model the relationship between each tag and the tokens of the diagnostic captions it is associated with in the training data. It is the first guided decoding method specifically developed for DC, as well as the first *data-driven* method that uses image tags to guide the generation of image captions.

*Corresponding author.

DMMCS is applicable to any encoder-decoder DC system, as it can be integrated in the decoding process. For experimental purposes, we train four DC systems, ranging from a generic CNN-RNN image-to-text 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 investigate the impact of DMMCS on few-shot captioning scenarios. We evaluate the performance of all models on two medical datasets, ImageCLEFmedical 2023 (Rückert et al., 2023) and MIMIC-CXR (Johnson et al., 2019). We use two evaluation measures, BLEU and BLEURT, comparing the results with and without DMMCS, demonstrating the effectiveness of the proposed algorithm. As one would expect, the performance boost provided by DMMCS is larger when using gold tags, but we show that DMMCS is also beneficial with noisy tags predicted by medical image classifiers.

Our main contributions are: (i) We introduce DMMCS, a data-driven guided decoding method for DC, which leverages medical tags of the input images to improve the generated captions. (ii) We incorporate DMMCS in four DC systems covering a range of learning scenarios, including fine-tuning and few-shot learning. (iii) We demonstrate that DMMCS significantly enhances performance across all four models in most cases, even when using noisy predicted tags.

2 Related Work

Substantial research has been dedicated to Controllable Text Generation and guided decoding strategies (Prabhumoye et al., 2020). Standard decoding methods, such as greedy or standard beam 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. Most 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 score is 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 related method for semantic guided decoding, which comprised the starting point of our research, is *Contrastive Search* (Su et al., 2022). It attempts to tackle the problem of text degeneration, where models produce unnatural and repetitive text. It imposes a degeneration penalty at each decoding step to guide the model's output towards more natural and less repetitive text sequences. The penalty for a candidate token u at decoding step t is:

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

where x_j are the preceding tokens of the incomplete generated sequence x, $h\left(\cdot\right)$ calculates word embeddings, and $sim(\cdot, \cdot)$ denotes cosine similarity. The degeneration penalty D_u is the maximum cosine similarity between the word embeddings of the candidate token u and the preceding sequence $x_{<t}$. It is subtracted from the score that the decoder would otherwise assign to u, thus guiding the decoder to generate less repetitive text.

3 The Proposed DMMCS Method

Our proposed method introduces a tag-guided decoding strategy for DC systems. It aims to guide the model to select words that appropriately express the tags (medical concepts) associated with

the input image.[1] 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 probably inaccurate.

As a first exploration, we calculated the FastText word embeddings (Bojanowski et al., 2017) of all the tags and all the tokens of the gold captions in the training set of the ImageCLEF 2023 dataset (Rückert et al., 2023). Tags consisting of multiple tokens were represented by the centroid of the tokens' embeddings. We then investigated the relationship between each tag and the gold captions it was associated with (the gold captions of training images tagged with the particular tag). Figure 1 presents a heatmap that visualizes the relationship between the tokens of a caption s (x-axis) and its corresponding set of tags T (y-axis). Each heatmap cell 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,

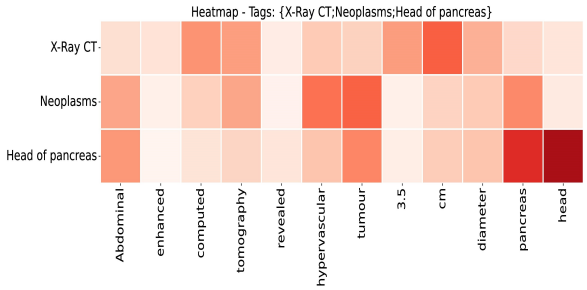

Figure 1: Heatmap visualizing the cosine embedding similarities between the tokens of a ground truth caption (x-axis) and its associated biomedical tags (y-axis).

for instance, the lower right cells show that tokens s_{11} and s_{12} ("pancreas head") of the caption have a very high cosine similarity with tag t_3 ("Head of pancreas"). Indeed, t_3 is almost explicitly expressed in s (almost the same words), while the other two tags are expressed more implicitly (with different words, e.g., "Neoplasms" vs. "tumour", "X-Ray CT" vs. "computed tomography"). We define the similarity between a tag t and a caption s as the *maximum cosine similarity (MCS)* between the (centroid) word embedding of t and the word embedding of each token in s, i.e.,

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

[1] In our experiments, the input is always a single image (and its tags), but DMMCS also applies to multi-image inputs.

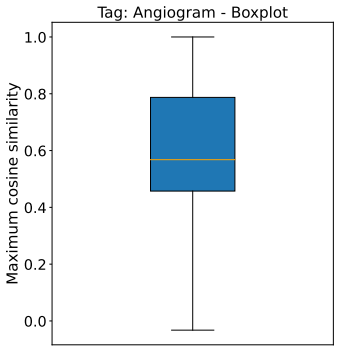

Figure 2: Distribution $R(t, s)$ for the tag "Angiogram". Interquartile range (IQR) shown as blue box. The coral line is the median, denoted $MMCS(t, S)$.

A high $MCS(t, s)$ between a tag t and a caption s indicates a significant presence of the tag's meaning in the caption. Next, we investigated the relationship between each tag t and the set S of all the gold captions tag t is associated with in the training data (gold captions of images tagged with t). For each tag t and its associated captions S, we compute the distribution $R(t, S)$ of its corresponding MCS scores, as the set:

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

Figure 2 illustrates the distribution $R(t, S)$ of the tag "Angiogram", using the gold captions of the images associated with this particular tag in the ground truth of the ImageCLEF 2023 training set. The blue box represents the interquartile range (IQR) of the distribution, while the coral line denotes the median value, denoted $MMCS(t)$:

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

We repeated the calculation of the distribution $R(t, S)$ for every tag t and the set S of gold captions associated with t in the training set of ImageCLEF 2023. Figure 3 shows the interquartile range (IQR, black vertical lines) and $MMCS(t, S)$ (median, coral line) of the distribution $R(t, S)$ for each tag t. In other words, Figure 3 contains many box-plots, like the one of Figure 2, side by side. Intuitively, Figure 3 shows how strongly each tag t is expressed in the gold captions associated with it. The side-by-side box-plots of Figure 3 are sorted by ascending $MMCS(t, S)$ (coral), in order to highlight the observation that the tags of the ImageCLEFmedical 2023 dataset are not expressed equally strongly in the ground truth captions. For instance, we observe a variation in $MMCS(t, S)$ values (coral line), 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 expressed more implicitly in the diagnostic captions they are associated with, while the latter are explicitly mentioned. Some tags conveying information that may be trivial to a clinician (e.g., that the image is an X-Ray) may actually not be expressed at all (not even implicitly).

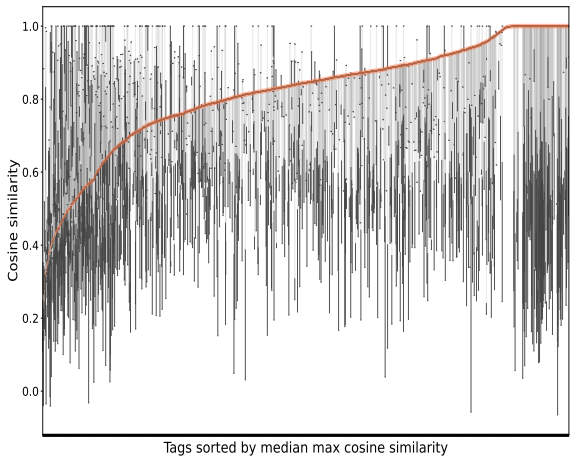

Figure 3: $MMCS(t, S)$ (coral line) and IQR (vertical lines) per tag t, sorted by ascending $MMCS(t, S)$.

Considering these findings, we propose a new decoding penalty that aims to improve the generated diagnostic captions by integrating information provided by the image's medical tags. The tags are in practice predicted by a medical image tagger, but we experiment with both predicted and gold (oracle) tags. The penalty encourages the decoder to select words that express more or less explicitly (or not at all) the tags of the image. The target level of explicitness of each tag is determined by its $MMCS(t, S)$ score, which is computed on the training captions. Tags with larger $MMCS(t, S)$ should be explicitly mentioned, while tags with lower $MMCS(t, S)$ should be expressed less explicitly (or not at all).

During inference, if an image is associated with a single tag t, we calculate the $MCS(t, s)$ (Eq. 2) between the tag t and the tokens of each candidate (possibly still incomplete) caption s being considered by the beam search decoder. The penalty is the squared difference between the computed $MCS(t, s)$, which shows how strongly the tag is expressed in the candidate caption, and the tag's $MMCS(t, S)$, which shows how strongly the tag is expressed (median value) in the ground-truth training captions it is associated with. If an image is associated with multiple tags, then a separate penalty is calculated for each tag, as above, and the total penalty is the sum of the penalties divided by the number of associated tags. Formally, given a (possibly incomplete) caption s and a set of image tags T to be expressed, the penalty is calculated as:

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

Intuitively, the goal is to generate a caption that expresses each tag of the image as strongly as the training captions associated with the tag.

At each decoding step, each candidate (possibly incomplete) caption s considered by the beam search decoder is scored as follows:

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

where T is the set of tags the input image is associated with. D_{score} and α are explained below.

D_{score}: This is the score the decoder assigns to each candidate caption, i.e., the sum of the (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 conducted experiments with both types of architectures in order to test the effectiveness of the DMMCS-based decoding. D_{score} is also normalized in $[0, 1]$, using min-max scaling, in order to align with the score range of $DMMCS_p$. As the goal is to minimize the overall score, Eq. 6 uses $1 - D_{score}$.

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

α: This hyper-parameter controls the effect of the two terms in the overall $DMMCS(s)$ score. The larger the value of α, the more significant the influence of the penalty $DMMCS_p$ on the overall score, and vice-versa. When $\alpha = 0$, standard beam search is applied. The value of α is tuned by experimenting with several values on the validation set.

In summary, at each decoding step the $DMMCS(s)$ score (Eq. 6) is calculated for each candidate sequence s of the beam search. The score combines $DMMCS_p$ (Eq. 5) and D_{score} (Eq. 7). The former measures how well the input image's tags are expressed in the candidate (possibly incomplete) caption, while the latter denotes the score

assigned by the decoder to the candidate caption. The n sequences with the best *DMMCS*(s) scores are selected and expanded at the next decoding step. Algorithm 1 presents our proposed guided decoding method in pseudo-code form.

Algorithm 1 DMMCS algorithm - Pseudocode

T, S, I ▷ the dataset tags, captions, images
$i \in I$ ▷ an image in question
$\text{tags}_i^p, \text{tags}_i^g \subseteq T$ ▷ predicted, gold tags of i

Phase 1 - Training Statistics

for $t \in T$ **do**
  $S' \leftarrow \{s_i \in S : t \in \text{tags}_i^g\}$ ▷ captions per t
  $\text{MMCS}(t) \leftarrow \text{median}\{\text{MCS}(t, s) | s \in S'\}$ ▷ Eqs. 2–4
end for

Phase 2 - Inference

for each decoding step of the beam search **do**
  for each beam search sequence s **do**
    get D_{score} ▷ decoder's score of s
    ▷ Calculate DMMCS penalty (Eq. 5)
    $T' \leftarrow \text{tags}_i^p$ ▷ predicted tags of i
    $\text{DMMCS}_p \leftarrow \frac{1}{|T'|} \sum_{t \in T'} (\text{MCS}(t, s) - \text{MMCS}(t))^2$
    ▷ Calculate DMMCS score (Eq. 6)
    $\text{DMMCS}_s \leftarrow \alpha \cdot \text{DMMCS}_p + (1 - \alpha) \cdot (1 - D_{score})$
    ▷ Set updated beam scores
    $\text{beamScores}[s] \leftarrow \text{DMMCS}_s$
  end for
end for

4 Experiments

For experimental purposes, we employed DMMCS decoding on four DC models, spanning diverse architectures and learning techniques, including long-established generic image captioning models and more advanced state-of-the-art ones. Moreover, we evaluated DMMCS using two datasets: the dataset of ImageCLEFmedical 2023 (Rückert et al., 2023), and a subset of the MIMIC-CXR dataset (Johnson et al., 2019). The performance of each model is evaluated using two measures, briefly discussed in §4.3. We demonstrate that DMMCS leads to substantial performance improvements in most cases.

4.1 Datasets

ImageCLEFmedical 2023 Dataset: This 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 gold tags and gold diagnostic captions. The maximum number of words in a single caption is 315, the minimum is 1, while the average is 16.04. The dataset contains 2,125 distinct tags, and the images are from multiple medical modalities, such as X-Ray, CT, MRI. No ground truth captions for the test set are available. We split the provided data into three subsets, following a 75%-10%-15% split, holding out a test subset for evaluation purposes. Thus, we employed 53, 516 images as our training data, 7, 135 images as our validation set, while the remaining 10, 704 images comprised our held-out test set.

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 gold medical tags and a gold diagnostic caption. The maximum number of words in a single caption is 1, 014, the minimum is 2, while the average is 137.8. There are 14 distinct biomedical tags in total, generated by the CheXpert labeler system (Irvin et al., 2019b). We only considered examinations containing a single radiology image per patient, because the DC systems we used could only receive a single image as input. Thus, we used 32,637 images, along with their gold tags and captions. We randomly split them into training, validation and test sets, containing 25,301, 2,752 and 4,584 samples, respectively.

4.2 DC Models

Show & Tell: This is a basic generic image captioning model (Vinyals et al., 2015), consisting of a CNN image encoder followed by an RNN text decoder (no attention used).[2] The CNN extracts image features that are then fed to the RNN decoder, which generates the diagnostic caption word by word, based on the image features.

ViT-GPT2: This is also based on the encoder-decoder approach, but now both components are based on Transformers (Vaswani et al., 2017). Vi-

[2]Show, Attend & Tell (Xu et al., 2015) with attention performed worse in DC in preliminary experiments.

sion 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 HuggingFace checkpoint[3] for a joint ViT-GPT2 encoder-decoder pipeline, and were further pre-trained on generic image-caption pairs. We then fine-tuned the model on the two employed DC datasets.

InstructBLIP: This is vision-language instruction-tuned model (Dai et al., 2023). Such models are designed to swiftly adapt to new tasks based on specific instructions. Hence, their performance strongly depends on the provided instructions (presented in Appendix F).

Flamingo: This is a few-shot (in-context learning) generic image captioning system (Alayrac et al., 2022). Flamingo can generate a diagnostic caption based on a few demonstrative examples of image-caption pairs provided as a multi-modal prompt. Checkpoints of the original Flamingo architecture are not publicly available, hence we employed OpenFlamingo, an open-source implementation that obtains around 80-89% of the original Flamingo's performance (Awadalla et al., 2023).

4.3 Evaluation Measures

BLEU, BLEURT: We use BLEU (Papineni et al., 2002) and the more recent (BERT-based) BLEURT (Sellam et al., 2020) as main evaluation measures. Both are frequently used in text generation, including generic image-to-text and DC.[4]

Clinical Accuracy: Previous work has shown measures like BLEU and BLEURT may not adequately capture clinical correctness in DC (Pavlopoulos et al., 2021). Hence, we also measure *clinical accuracy* (CA) by comparing medical concepts extracted from the generated diagnostic captions (as silver labels by a multi-label text classifier) to those extracted from the corresponding gold captions. To compute CA, we follow previous work (Liu et al., 2019) that employs CheXBERT (Irvin et al., 2019a) to determine whether or not each one of the 14 thoracic ailments (treated as classes) are mentioned in a given diagnostic caption. Given a caption (generated or gold), CheXBERT produces "present", "negative", "unsure" or "blank", for each one of the 14 ailments. We treat "blank" as "negative". When the "unsure" label is predicted, we change the label to either "present" or "negative" with equal probability, as in the work of Liu et al. (2019). CA is defined as:

$$CA = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{n}\sum_{i=1}^{n} 1\{y_R[i,j] = y_P[i,j]\},$$

where n is the number of classes, m is the number of caption pairs (gold-generated) being compared, and y_R, y_P denote predictions from reference (gold) and predicted (generated) captions, respectively.

4.4 Baselines

Beam search decoding: The first baseline we compare DMMCS against is standard beam search.

Constrained beam search: The second baseline is a constrained beam search decoding method (Anderson et al., 2017), where decoding is guided by a set of manually defined constraints, aiming to enforce specific lexical requirements in the generated captions. We experimented with two versions of constrained beam search, namely strict and disjunctive. The former (denoted $\forall$) ensures that all specified keywords are present in the generated captions. The latter (denoted $\exists$) only enforces the inclusion of at least one of the given tags in the generated captions. For both versions, we used the HuggingFace implementation of constrained beam search,[5] which is based on a plethora of guided decoding methods (Anderson et al., 2017; Post and Vilar, 2018; Hu et al., 2019; Li et al., 2021).

4.5 Experimental Results

We repeat each experiment for three random non-intersecting subsets of the test set, with each subset containing $1,000$ images. We report the average (over the the three test subsets) score for each evaluation measure and the standard deviation, offering insights into the stability of each model's performance across different test subsets. To obtain the medical tags of each image, we use a medical image tagger trained on the training subset of each dataset. Specifically, we employ the top-performing encoder of the ImageCLEFmedical 2023 campaign (Ionescu et al., 2023; Rückert et al., 2023), namely a DenseNet-121 instance (Huang et al., 2017), initially pre-trained on ImageNet (Deng et al., 2009). We fine-tuned it (separately per DC dataset) on the training images and corresponding gold tags of the two DC datasets we

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

[4] We also report ROUGE scores in Appendix B.

[5] https://huggingface.co/blog/constrained-beam-search

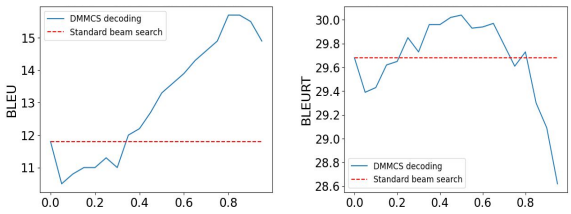

Figure 4: InstructBLIP's performance with DMMCS decoding for various α values (horizontally) on ImageCLEF medical 2023, when using BLEU (left) or BLEURT (right) to measure performance.

use (Section 4.1). We also report scores using the gold (oracle) tags of the test images (Appendix B).

Tuning α: For each model, dataset, and evaluation measure, we tuned the α factor of DMMCS (Eq. 6) by trying values from 0.05 to 0.95 and keeping the value with the best development score (measured with the particular evaluation measure). As an example, Fig. 4 shows the performance of InstructBLIP on ImageCLEFmedical 2023, in terms of BLEU (left) or BLEURT (right), when using DMMCS (continuous lines) or standard beam search decoding (dotted horizontal lines), as a function of α. For α values in $[0.4, 0.8]$, DMMCS improves the model's performance, while smaller and larger values deteriorate it, comparing to standard beam search.

BLEU, BLEURT results: Table 1 reports the performance of each model on the two datasets, averaging over the three test subsets. For each evaluation measure (BLEU, BLEURT), four scores are provided: one for each baseline method (standard beam search, strict and disjunctive constrained beam search), as well as one for our proposed DMMCS method. The scores for DMMCS are obtained with the best α for each model, dataset, measure, using development data, as discussed above. Tags predicted by a medical image tagger are used. We observe that DMMCS always outperforms both standard (BS) and strict constrained ($\forall$) beam search. Moreover, DMMCS is on par with the disjunctive constrained ($\exists$) beam search method, outperforming it in most cases. We show the number of "wins" of each decoding method in the last row of Table 1. We also experimented using gold (oracle) tags instead of predicted tags; the results, discussed in Appendix B, show that DMMCS again improves performance, comparing to standard beam search.

Clinical accuracy results: In Table 2, we present the best performing method across models and datasets in terms of clinical accuracy (CA). When InstructBLIP is used as the backbone model, DMMCS is the best mechanism across datasets. For the rest of the models, however, there is no clear winner, although ConBS$\forall$ is the best in Mimic (see Appendix I). We note that CA is based on silver labels (automatically generated), for a limited number of classes, and only on captions regarding chest (i.e., less than 20%). Consequently, the reliability of these results may be compromised to some extent. Future work should focus on the development of better medical image taggers that could improve automated clinical evaluation (e.g., making it applicable not only to chest images).

Fluency results: In the last two rows of Table 2 we present the results based on fluency (using Perplexity, see Appendix H) instead of CA. Our proposed method is the best in both InstructBLIP and ViT-GPT2, and the best across models for ImageCLEFmedical 2023.

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 show a sample from this analysis 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, unlike beam search decoding which incorrectly stated that the examination type showed a "left ventricular outflow tract". However, it is crucial to note 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, and in practice they would be checked and improved by medical experts.

Varying Sentence Order Analysis: The sentence order in medical reports varies (e.g., the same diagnosis may appear in different reports, using the same sentences yet reordered). In light of this

ImageCLEFmedical 2023 Dataset - With Predicted Tags								
	BLEU				BLEURT			
	BS	*ConBS* ∀	*ConBS* ∃	*DMMCS*	*BS*	*ConBS* ∀	*ConBS* ∃	*DMMCS*
Show & Tell	20.61 (0.33)	20.52 (0.39)	21.21 (0.38)	**21.27** (0.35)	29.99 (0.14)	30.03 (0.17)	30.39 (0.13)	**30.47** (0.08)
ViT-GPT2	15.34 (0.09)	15.75 (0.12)	16.29 (0.08)	**16.31** (0.08)	26.50 (0.09)	26.31 (0.10)	26.92 (0.14)	**27.01** (0.16)
InstructBLIP	11.81 (0.09)	15.89 (0.08)	**16.14** (0.13)	15.93 (0.11)	29.68 (0.26)	29.71 (0.12)	30.08 (0.14)	**30.10** (0.15)
Flamingo	15.34 (0.11)	15.81 (0.13)	**15.92** (0.06)	15.47 (0.09)	28.49 (0.19)	30.11 (0.21)	30.67 (0.19)	**31.34** (0.16)
MIMIC-CXR Dataset - With Predicted Tags								
	BLEU				BLEURT			
	BS	*ConBS* ∀	*ConBS* ∃	*DMMCS*	*BS*	*ConBS* ∀	*ConBS* ∃	*DMMCS*
Show & Tell	11.77 (0.18)	12.14 (0.22)	13.38 (0.17)	**14.13** (0.24)	29.08 (0.27)	29.04 (0.29)	29.21 (0.31)	**29.49** (0.29)
ViT-GPT2	13.55 (0.26)	12.91 (0.19)	13.67 (0.23)	**14.78** (0.21)	23.52 (0.34)	24.87 (0.37)	**24.98** (0.30)	24.37 (0.29)
InstructBLIP	12.76 (0.20)	12.09 (0.22)	12.27 (0.17)	**13.32** (0.19)	24.65 (0.24)	26.28 (0.27)	**26.43** (0.26)	25.56 (0.25)
Flamingo	12.78 (0.11)	13.07 (0.13)	**13.39** (0.11)	13.26 (0.16)	29.14 (0.22)	28.86 (0.21)	29.58 (0.27)	**29.81** (0.24)
Wins	0	0	3	**5**	0	0	2	**6**

Table 1: The performance of each model on both datasets, measured by BLEU or BLEURT. BS and ConBS denote beam search and constrained beam search decoding, respectively. ∀ and ∃ indicate whether ConBS is strict (all image tags must be expressed) or disjunctive (only one suffices), respectively. DMMCS is the new proposed decoding method. For the latter, tags predicted by a medical image tagger are used. We also report, along the evaluation metric score, the variance between the results of the three test subsets.

	Show & Tell	**ViT-GPT2**	**InstructBLIP**	**Flamingo**
ImageCLEFmedical 2023	BS	ConBS ∃	DMMCS	BS
MIMIC-CXR	ConBS ∀	ConBS ∀	DMMCS	ConBS ∀
ImageCLEFmedical 2023	DMMCS	DMMCS	DMMCS	DMMCS
MIMIC-CXR	ConBS ∃	DMMCS	DMMCS	BS

Table 2: The first two rows present the most clinically accurate guided decoding method per model per dataset (more details in Appendix I). The last two rows present the same leaderboard, but use fluency (measured as Perplexity, Appendix H) instead of clinical accuracy.

observation, we conducted a sentence-level analysis to investigate the potential influence of sentence order on the quality of the generated captions. We considered pairs of gold and generated captions consisting of the same number of sentences. We then created sentence pairs using sentences at the same position in the gold and generated captions. We evaluated each pair in terms of BLEU and BLEURT and then averaged the sentence-pair scores across the test set. If the order of the generated sentences is often wrong compared to the corresponding gold ones, we should notice a substantial difference relative to the original scores, which do not penalize generating reasonably correct sentences but in the wrong order. Our analysis, however, revealed that the scores of sentence pairs remained consistent with the original ones (within the reported standard deviation), ultimately indicating that this issue did not affect our experiments.

Cases where performance declines: As shown in Tables 1 and 2, the proposed algorithm generally improves performance on both Natural Language Generation (NLG) metrics and clinical accuracy. However, there are instances where the performance declines when using the proposed algorithm. We investigated these cases by manually examining them, but no specific pattern explaining the decrease in performance could be identified. We plan to investigate this issue further in the future, possibly requiring assistance from medical professionals.

Computational Overhead: It is important to measure the computational overhead associated with our proposed method. Implementing our method results in an additional time overhead, requiring approximately 25 to 27% more time to generate diagnostic captions compared to baseline methods such as standard beam search. In the case of InstructBLIP, this translates to an increase of around 13 to

15 minutes in processing time per $1,000$ captions in our experiments. Furthermore, there is an additional computational overhead in terms of memory usage, which increases by approximately 5% compared to baseline approaches (Section 4.4). While these overheads are important considerations, the improved performance, in terms of NLG metrics and clinical accuracy, of the generated captions justify the investment in time and computational resources.

Per-Modality Results The ImageCLEFmedical 2023 dataset consists of four primary medical modalities; namely X-Ray, CT, MRI and Ultrasonography. We explored the individual per-modality results of both the standard beam search and the proposed decoding method in order to identify any modality-specific features that might require attention. We observed that the proposed algorithm consistently outperforms the standard beam search across both metrics and nearly all modalities and models. No modality-specific peculiarities were noted, so we did not deem it necessary to implement any case-specific data handling procedures. Detailed per-modality evaluations are provided in Table 4 (Appendix G).

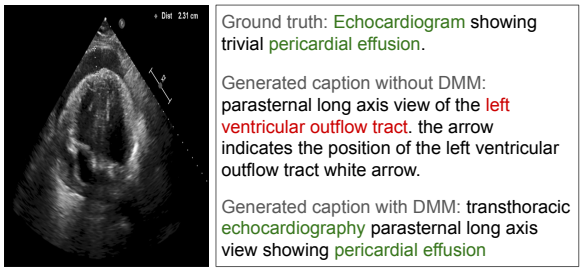

Figure 5: A sample from an exploratory qualitative analysis of captions generated by ViT-GPT2, using standard beam search vs. DMMCS-enhanced decoding.

5 Discussion

Performance with gold tags

We also experimented with the ground-truth (instead of the predicted) tags per image (see also Table 3 in the Appendix). As one would expect, DMMCS-enhanced models perform better with gold, compared to predicted tags. However, the performance of DMMCS-enhanced models with predicted tags (Table 1) is relatively close to the corresponding performance with ground-truth tags (Table 3), showing the robustness of DMMCS to noisy tags.

Guidance balancing

We also experimented by balancing the two basic components of DMMCS (Eq. 6), the decoder's scores and the proposed data-driven penalty. By introducing a dynamically computed weight called Histogram Divergence (HD), we adjusted the contribution of the two components, so that when DMMCS could not be trusted, more weight would be assigned to the decoder's score (more details in Appendix E). We observe that this balancing results in steadier performance across α values, though it comes at the cost of lower overall performance compared to when HD is not considered. In other words, if tuning is not an option, dynamically balancing the two components of DMMCS could lead to the selection of a better optimal α.

6 Conclusion

In this work we introduced the DMMCS data-driven guided decoding method that enhances the performance 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 DC systems with diverse architectures and on two datasets, namely ImageCLEFmedical 2023 and a subset of MIMIC-CXR. Subsequently, we compared the generated captions with those obtained using a typical beam search decoding approach, employing widely used evaluation metrics, namely BLEU and BLEURT. Our results demonstrate that using the proposed DMMCS mechanism during decoding consistently outperforms the typical beam search approach across almost all models and datasets for most of the metrics.

In future work, we plan to experiment with more 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. A final direction for future work concerns the use of contextual representations, in order to enhance the quality of the embeddings used when computing the penalty of DMMCS (Eq. 5).

7 Limitations

Although we used all the publicly available medical datasets we could obtain, the experimental results are limited to the specific conditions, regions, and

language (English) these datasets concern. However, 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.

8 Ethical Considerations

Assisting 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.

Acknowledgements

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program.

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.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2017. Guided open vocabulary image captioning with constrained beam search. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 936–945, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Shuang Bai and Shan An. 2018. A survey on automatic image caption generation. *Neurocomputing*, 311:291–304.

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.

J. Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, and Benjamin Van Durme. 2019. Improved lexically constrained decoding for translation and monolingual rewriting. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 839–850, Minneapolis, Minnesota. 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.

J. A. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. L. Ball, K. S. Shpanskaya, J. Seekins, D. Andrew Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and A. Ng. 2019a. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI Conference on Artificial Intelligence*.

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. 2019b. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. AAAI'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.

Alistair Johnson, Tom Pollard, Lu Shen, Li Wei Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Celi, and Roger Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035.

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.

Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 1537–1547.

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

Zhongyang Li, Xiao Ding, Ting Liu, J. Edward Hu, and Benjamin Van Durme. 2021. Guided generation of cause and effect. IJCAI'20.

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.

Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. 2019. Clinically accurate chest x-ray report generation. In *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pages 249–269. PMLR.

M. Y. Lu, R. J. Chen, and F. Mahmood. 2020. Semi-supervised breast cancer histology classification using deep multiple instance learning and contrast predictive coding (conference presentation). In *Medical imaging 2020: digital pathology*. SPIE.

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, 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*.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Pranav Rajpurkar, Jeremy A. Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Yi Ding, Aarti Bagul, C. Langlotz, Katie S. Shpanskaya, Matthew P. Lungren, and A. Ng. 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *ArXiv*.

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.

S. Varges, H. Bieler, M. Stede, L. C. Faulstich, K. Irsig, and M. Atalla. 2012. Semscribe: Natural language generation for medical reports. In *International Conference on Language Resources and Evaluation*.

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.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France.

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.

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.

Appendix

A Background on DC

Various DC approaches have been explored in previous work, ranging from earlier ontology and rule-based systems (Varges et al., 2012) to modern encoder-decoder architectures. Notably, the best-performing techniques in generic image captioning (Bai and An, 2018) are not always the most efficient for DC tasks. Unlike generic image captioning models, DC systems face challenges in accurately describing medical images based solely on visual content. For instance, conventional retrieval methods, despite their simplicity, have shown promising results in DC tasks by leveraging captions from similar archived exams (Pavlopoulos et al., 2021). Even simple models like 1-NN, which retrieve and use the caption from the visually nearest image, can outperform more complex systems, such as Transformer-based or encoder-decoder architectures (Liu et al., 2019). However, despite the noteworthy performance of retrieval methods, recent advancements in the field of deep learning have established the encoder-decoder framework as the predominant approach in the DC domain. CNNs or Vision Transformers (Dosovitskiy et al., 2021) are commonly employed for image encoding, while RNNs or Transformer-based LMs are selected for the report generation process. Recent advancements in diagnostic captioning have also incorporated attention mechanisms, visual attention, and reinforcement learning techniques into the encoder-decoder framework (Li et al., 2018).

B DMMCS based on ground-truth tags

In this section, we report the performance of the four models on both datasets using the *ground-truth* tags for each image (Table 3), instead of predicted tags (Table 1). As shown in the main section of the paper, employing the proposed algorithm enhances model performance compared to standard beam search. Furthermore, employing ground-truth tags, instead of tags predicted by a medical image classifier, results in superior model performance. The extent of this performance difference depends on the accuracy of the classifier. Less accurate classifiers generate noisier tags, which subsequently affect the generated captions. Conversely, more accurate classifiers provide guidance for the model to include relevant concepts, as well as words closely aligned with the medical examination's findings. In our experiments, we observed that the performance of the models utilizing predicted tags (Table 1) was relatively close to the ones employing ground-truth tags (Table 3). This indicates that despite the use of predicted tags, the model's performance remained competitive, indicating the robustness of the proposed decoding algorithm.

C ROUGE scores

In the main part of this paper, we primarily focused on a precision-based measure (i.e., BLEU (Papineni et al., 2002)) and a learned evaluation metric (i.e., BLEURT (Sellam et al., 2020)) that relies on precision-driven metrics during training. BLEU evaluates the precision of n-grams in the generated text compared to reference texts, while BLEURT is a learned metric that leverages transformer-based embeddings and is trained on human judgment data. It incorporates BLEU in order to enhance its evaluation capabilities. Nevertheless, recall is also important, as it measures the model's ability to capture all relevant details from the reference captions. Unlike precision, which emphasizes the correctness of the generated information, recall ensures that the model doesn't miss any pertinent details. However, clinicians face time constraints and cannot feasibly process overly long or potentially inaccurate information. Therefore, while recall is crucial for ensuring thoroughness, it must be balanced with

A. ImageCLEFmedical 2023 Dataset - Ground-Truth Tags						
	BLEU		**ROUGE**		**BLEURT**	
	BS	*DMMCS*	*BS*	*DMMCS*	*BS*	*DMMCS*
Show & Tell	20.61	**21.48**	22.00	**23.01**	29.99	**30.53**
ViT-GPT2	15.34	**16.72**	16.72	**17.05**	26.50	**27.69**
InstructBLIP	11.81	**19.95**	20.98	**21.17**	29.68	**30.65**
Flamingo	15.34	**15.83**	15.98	**16.11**	28.49	**31.47**
B. MIMIC-CXR Dataset - Ground-Truth Tags						
	BLEU		**ROUGE**		**BLEURT**	
	BS	*DMMCS*	*BS*	*DMMCS*	*BS*	*DMMCS*
Show & Tell	11.77	**14.29**	**17.18**	17.10	29.08	**29.65**
ViT-GPT2	13.55	**15.01**	21.37	**21.74**	23.52	**24.43**
InstructBLIP	12.76	**13.39**	14.35	**16.27**	24.65	**25.61**
Flamingo	12.78	**13.99**	13.68	**13.86**	29.14	**29.87**

Table 3: The performance of each model on both datasets evaluated across the three metrics. *BS* and *DMMCS* denote beam search and DMMCS-based decoding respectively. For the DMMCS-based decoding, ground-truth tags provided in the training set are used.

the practical considerations of clinical workflow. Our main focus remains on precision, ensuring that the information provided is concise, relevant, and accurate. Nevertheless, the model's performance in a recall-based metric, specifically ROUGE (Lin, 2004), is also included in this section (Table 3). This approach allows us to provide a more comprehensive evaluation while maintaining a focus on the precision needed for practical clinical use.

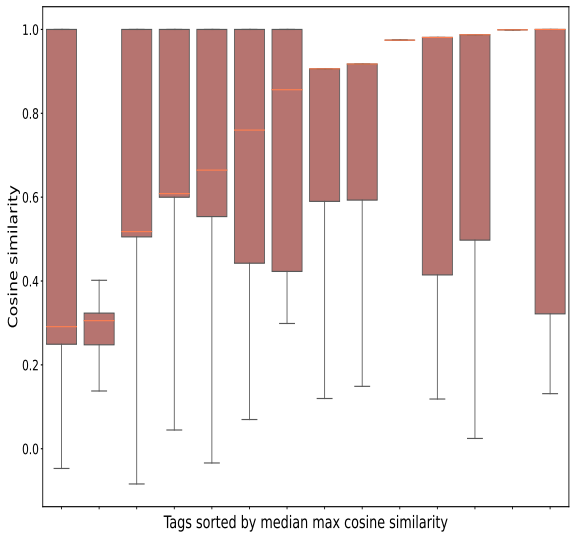

Figure 6: Multiple boxplots plotted side-by-side. Each plot visualizes the *MCS* distribution between a tag and its associated training captions in the MIMIC-CXR dataset. The boxplots are sorted based on their median value (coral line), which denotes the tag's median maximum cosine similarity (MMCS) value.

D MIMIC-CXR MCS distributions

In this part, we also provide a figure with multiple interquartile plots for the employed subset of the MIMIC-CXR (Johnson et al., 2019) dataset, 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.

E Histogram Divergence (HD)

As part of further exploration, we experimented with adding an additional term in our method's guided decoding scoring function. Histogram Divergence (HD) is a dynamically computed parameter that serves as a weighting factor of the two components, $DMMCS_p$ and $1 - D_{score}$. In detail, at each decoding step, we consider a set of generated candidate sequences G and a histogram of the MCS distribution $R(t, G)$ (Eq. 3) 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 are generated. In addition, a similar histogram has been pre-calculated for each tag on its associated ground-truth captions with respect to the training data (see Section 3). The rationale behind HD is that in candidate sequences with similar *MCS* distributions (one for each tag) to those computed on the training data the $DMMCS_p$ factor should be assigned a larger weight compared to $1 - D_{score}$ during that decoding step. In contrast, if the distributions do not match, the conventional 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, while it is an appropriate metric in the specified context since 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:

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

HD is integrated in Eq. 6 as an additional weighting factor:

$$\begin{aligned} DMMCS(s) = \alpha \cdot (1 - HD) \cdot DMMCS_p(T, s) \\ + (1 - \alpha) \cdot HD \cdot (1 - D_{score}) \end{aligned} \quad (10)$$

As can be seen in Fig. 7, both blue (with HD) and green (without HD) lines surpass the red dashed line of the beam search baseline for various α values when using the (oracle) gold tags. The green line achieves a higher peak compared to the blue, while the blue line has lower standard deviation. Therefore, HD improves robustness across α values at the cost of (maximum) performance.

F Instructions

InstructBLIP (Dai et al., 2023) operates as a Vision-Language instruction-tuning model, which can adapt rapidly to new tasks based on specific instructions provided in its query prompt during training. Consequently, its performance depends largely on the quality of the instruction it is prompted with. While identifying the optimal instruction remains elusive, it is commonly agreed that concise and coherent instructions tend to yield better performance. However, due to the substantial memory demands of an InstructBLIP instance, conducting numerous experiments to find the optimal or near-optimal instruction can be very impractical. As a result, we conducted experiments using three different instruction prompts, which are presented in Table 5.

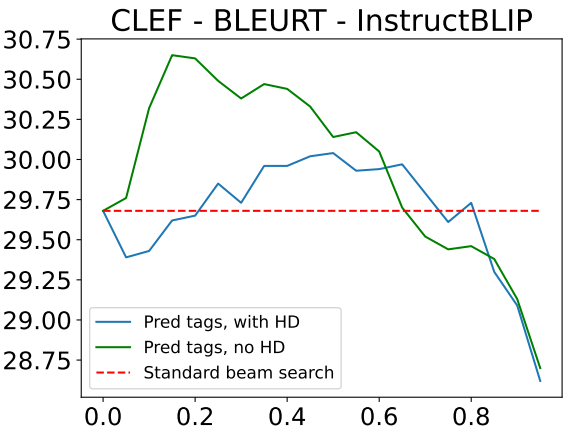

Figure 7: InstructBLIP's performance in BLEURT with (blue) and without (green) HD in ImageCLEFmedical 2023 for various α values (horizontally). The red line denotes the performance of the standard beam search baseline.

G Per-modality scores

In this section, we conducted a modality-specific evaluation in order to assess and highlight any notable differences in performance between different modalities, which is presented in Table 4. After our exploratory analysis, we observed that the ImageCLEFmedical 2023 dataset contains examinations originating from four main medical modalities, which are described in Thus, we split it into four subsets and evaluated the performance of the standard beam search and our proposed decoding method individually. In the rare case that an image belongs to more than one modality, we randomly select one of them to assign it. We observed that DMMCS outperforms beam search across almost all methods and modalities, remaining consistent with our findings on the entire dataset.

ImageCLEFmedical 2023 Dataset - Per modality Evaluation - With Predicted Tags								
	BLEU							
	X-Ray		**CT**		**MRI**		**Ultrasonography**	
	BS	*DMMCS*	*BS*	*DMMCS*	*BS*	*DMMCS*	*BS*	*DMMCS*
Show & Tell	20.73	**21.52**	20.71	**21.29**	20.54	**21.21**	20.46	**21.06**
ViT-GPT2	15.44	**16.62**	15.58	**16.21**	15.49	**15.87**	14.85	**16.54**
InstructBLIP	11.72	**15.92**	11.94	**16.07**	11.80	**16.01**	11.79	**15.72**
Flamingo	15.39	**15.63**	15.23	**15.67**	**15.41**	15.36	**15.25**	15.22
	BLEURT							
	X-Ray		**CT**		**MRI**		**Ultrasonography**	
	BS	*DMMCS*	*BS*	*DMMCS*	*BS*	*DMMCS*	*BS*	*DMMCS*
Show & Tell	30.13	**30.39**	29.96	**30.52**	30.03	**30.42**	29.84	**30.55**
ViT-GPT2	26.42	**27.17**	26.71	**26.82**	26.36	**26.94**	26.51	**27.11**
InstructBLIP	29.83	**30.13**	29.45	**30.01**	29.70	**30.27**	29.74	**29.99**
Flamingo	28.64	**31.33**	28.52	**31.61**	28.03	**31.31**	28.77	**31.11**

Table 4: The per-modality performance of each model on the ImageCLEFmedical 2023 dataset, measured by BLEU and BLEURT. BS denotes beam search, while DMMCS indicates the new proposed decoding method. For the latter, tags predicted by a medical image tagger are used.

InstructBLIP - Instruction Prompts	
1	"Describe the given radiology image."
2	"You are an experienced radiologist. You are being given radiology images along with a brief medical diagnosis. Generate a descriptive caption that highlights the location, nature and severity of the abnormality of the radiology image."
3	"You are a helpful medical assistant. Generate a diagnostic report based on the patient's radiology examinations."

Table 5: The three instructions that we defined in order to guide the InstructBLIP model throughout the DC task.

H Perplexity

We also provide perplexity scores (Table 6) obtained using ClinicalT5 to measure the fluency of each model.[6] ClinicalT5 is a biomedical version of T5 (Raffel et al., 2019), pre-trained on the MIMIC-III dataset (Johnson et al., 2016). We observe that our proposed method generates more fluent captions than the baseline decoding methods, as it achieves lower perplexity scores in most cases.

[6]We were granted access to ClinicalT5 through PhysioNet: https://www.physionet.org/content/clinical-t5/1.0.0/, Last accessed: 2024-06-05.

ImageCLEFmedical 2023 Dataset - Predicted Tags				
	Perplexity			
	BS	*ConBS* $\forall$	*ConBS* $\exists$	*DMMCS*
Show & Tell ($\times 10^4$)	18.38	31.12	18.75	**18.25**
ViT-GPT2 ($\times 10^4$)	17.70	26.96	20.30	**17.51**
InstructBLIP ($\times 10^4$)	16.47	16.61	22.12	**15.95**
Flamingo ($\times 10^4$)	20.79	23.69	21.02	**20.62**
MIMIC-CXR Dataset - Predicted Tags				
	Perplexity			
	BS	*ConBS* $\forall$	*ConBS* $\exists$	*DMMCS*
Show & Tell ($\times 10^6$)	3.05	3.03	**2.88**	3.01
ViT-GPT2 ($\times 10^7$)	3.01	3.40	3.83	**2.98**
InstructBLIP ($\times 10^8$)	1.39	1.26	1.35	**1.25**
Flamingo ($\times 10^7$)	**2.83**	3.31	3.85	2.90
Wins	1	0	1	**6**

Table 6: The perplexity of each model and decoding method on the ImageCLEFmedical 2023 and the MIMIC-CXR dataset computed using ClinicalT5.

I Clinical Accuracy - Extensive Results

Table 7 presents an extensive evaluation of the Clinical Accuracy measure across all models and decoding methods. We share more details on how we

calculate the CA between the gold and generated captions in Section 4.3.

ImageCLEFmedical 2023 - With Predicted Tags				
	Clinical Accuracy			
	BS	*ConBS* $\forall$	*ConBS* $\exists$	*DMMCS*
Show & Tell	**92.03**	90.82	91.73	91.38
ViT-GPT2	91.55	90.51	**91.77**	91.60
InstructBLIP	91.47	88.00	90.86	**91.73**
Flamingo	**91.29**	90.86	89.82	90.99
MIMIC-CXR Dataset - With Predicted Tags				
	Clinical Accuracy			
	BS	*ConBS* $\forall$	*ConBS* $\exists$	*DMMCS*
Show & Tell	83.87	**92.73**	84.51	84.80
ViT-GPT2	84.77	**91.78**	87.91	84.74
InstructBLIP	83.33	80.03	77.45	**84.19**
Flamingo	80.12	**88.82**	85.82	79.85
Wins	2	**3**	1	2

Table 7: The clinical accuracy of each model and decoding method on the ImageCLEFmedical 2023 and MIMIC-CXR datasets.