Factual Accuracy is not Enough: Planning Consistent Description Order for Radiology Report Generation

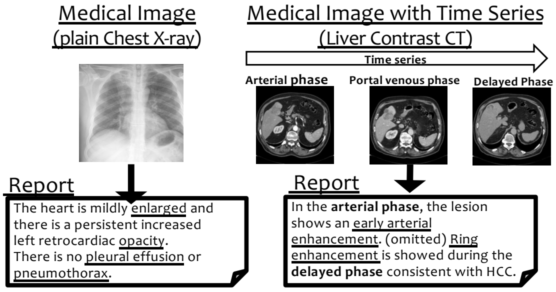
Toru Nishino[1] Yasuhide Miura[1] Tomoki Taniguchi[1] Tomoko Ohkuma[1]

Yuki Suzuki[2] Shoji Kido[2] Noriyuki Tomiyama[2]

[1]Fujifilm Corporation [2]Osaka University Graduate School of Medicine
toru.nishino@fujifilm.com

Abstract

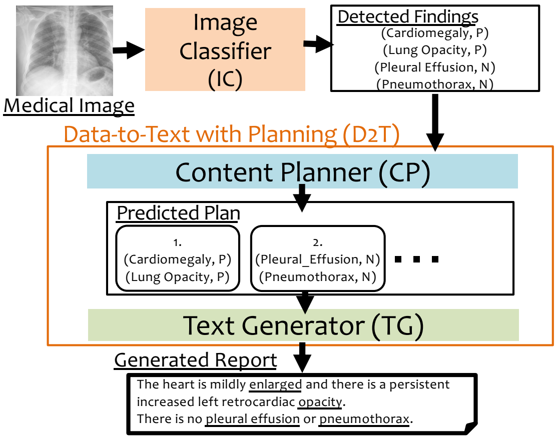
Radiology report generation systems have the potential to reduce the workload of radiologists by automatically describing the findings in medical images. To broaden the application of the report generation system, the system should generate reports that are not only factually accurate but also chronologically consistent, describing images that are presented in time order, that is, the correct order. We employ a planning-based radiology report generation system that generates the overall structure of reports as "plans" prior to generating reports that are accurate and consistent in order. Additionally, we propose a novel reinforcement learning and inference method, Coordinated Planning (CoPlan), that includes a content planner and a text generator to train and infer in a coordinated manner to alleviate the cascading of errors that are often inherent in planning-based models. We conducted experiments with single-phase diagnostic reports in which the factual accuracy is critical and multi-phase diagnostic reports in which the description order is critical. Our proposed CoPlan improves the content order score by 5.1 pt in time series critical scenarios and the clinical factual accuracy F-score by 9.1 pt in time series irrelevant scenarios, compared those of the baseline models without CoPlan.



The heart is mildly enlarged and there is a persistent increased left retrocardiac opacity. There is no pleural effusion or pneumothorax.

In the **arterial phase**, the lesion shows an early arterial enhancement. (omitted) Ring enhancement is showed during the **delayed phase** consistent with HCC.

Figure 1: Comparison of reports of single-phase diagnosis and multiphase diagnosis with time series.



The heart is mildly enlarged and there is a persistent increased left retrocardiac opacity. There is no pleural effusion or pneumothorax.

Figure 2: Overview of our planning-based report generation system.

1 Introduction

Radiologists regularly write qualitative radiology reports to accurately describe the recognized findings in medical images. Recently, we can observe two different approaches to radiology imaging: a *single-phase* and a *multiphase* imaging method. A single-phase diagnostic approach, as applied in a plain X-ray machine, scans only once, while many modern procedures, including liver contrast CT, use a multiphase diagnostic method, scanning sequentially in a period of several minutes. The time-dependent scans are labeled as *phases*. Reports of single-phase diagnoses are time series irrelevant, so the factual accuracy is the most critical quality consideration. However, reports of multiphase diagnoses are time series critical, so a time-variable transition of lesions in the image must be correctly described with consistent description order. For example ,in Figure 1, a liver contrast CT report is composed to reflect the order of time series, with an arterial phase followed by a delayed phase; on the contrary, the chest X-ray reports are composed with no regard to timing because there are no phases.

To broaden the application of report generation systems to multiphase diagnosis, the system should also become capable of generating reports that are written in consistent description order. Radiolo-

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 7123–7138
December 7-11, 2022 ©2022 Association for Computational Linguistics

gists intend to write high-quality radiology reports with consistent description order so that doctors can review and understand the radiology reports within a short time (European Society of Radiology (ESR), 2011). The existing studies (Monshi et al., 2020) can support single-phase diagnostics only, particularly in chest X-ray images. As mentioned in the limitation section of (Nguyen et al., 2021), contemporary studies face obstacles to generating reports with time series.

This study aims to design an automated radiology report generation system that generates factually accurate reports for time series irrelevant diagnoses and consistent description order reports for time series critical diagnostics. As shown in Figure 2, we employ a planning-based report generation system that consists of three modules: image classifier, content planner, and text generator. The content planner generates the "plan," which represents the content and description order of the reports, and then the text generator predicts accurate and consistent description order reports. Planning-based models generate more faithful sentences without hallucination than end-to-end models (Ferreira et al., 2019).

However, planning-based approaches have a critical disadvantage; they cascade errors in modules within the system. To solve the error cascading problem, PlanGen (Su et al., 2021) employs a reinforcement learning that encourages the generated output to adhere to the given content plan, and DYPLOC (Hua et al., 2021) uses multiple plan candidates in the content realization process to reflect the dynamic nature of plans.

We propose Coordinated Planning-based text generation (CoPlan), a novel unified framework that trains and conducts inferences on the content planner and text generator in coordination to generate more accurate reports in a consistent order. CoPlan checks cascaded errors in the final output of the system with a report evaluator to generate more appropriate plans for creating accurate reports. We employ two types of report evaluators for CoPlan, a fact-based evaluator and a description-order-based evaluator, to generate factually accurate reports for time series irrelevant scenarios and consistent description order reports for time series critical scenarios.

The contributions of this study are as follows:

- We present a planning-based report generation framework to generate radiology reports with a factually accurate and consistent description order to broaden the application of the radiology report generation system to the multiphase diagnostics applications.
- We propose **CoPlan**, which trains and conducts inference on the content planner and the text generator in a coordinated manner to generate correct and order consistent reports.

We evaluate our proposed CoPlan in both time-series irrelevant and critical scenarios. The datasets of multiple languages and modalities are used: the JLiverCT dataset in which the reports are written with time series description and the MIMIC-CXR dataset (Johnson et al., 2019) containing time irrelevant reports (i.e. only factual accuracy is critical). The results of the automatic and human evaluations show that our proposed method improves the accuracy of the dataset without times series and improves the consistency of the description order of the dataset with time series, compared to those of the models without CoPlan.

2 Related Works

Radiology Report Generation. Most radiology report generation studies (Monshi et al., 2020) proposed end-to-end systems that generate reports directly from images. However, they still cannot generate sufficiently accurate and consistent written order reports to replace the human radiologists, particularly in the time series critical scenario. Kurisinkel et al. (2021) proposed a two-stage model that predicts image representations for each sentence first and then decodes sentence-by-sentence to generate reports in a more consistent order. TS-MRGen (Nishino et al., 2020) consists of an image classifier that reads images and a data-to-text module to control the findings that should be described in the reports. Our proposed system combines the advantages of the two methods mentioned above; it generates accurate reports in a consistent order and enables the radiologists to control the generation process in the system.

Planning-based Text Generation. While current studies are primarily based on end-to-end neural text generation models, a neural planning-based approach that combines advantages of traditional pipeline text generation and neural text generation is widely researched (Tang et al., 2022). Ma et al. (2019); Moryossef et al. (2019) conducted a planning-based neural data-to-text research which comprises text planning and text realization modules to realize controllable and faithful text generation. However, the cascading of errors is a known problem in planning-based models. Errors occurring in the text planning module significantly affect the quality of the output of the text realization module. Shen et al. (2020) proposed an end-to-end trainable planning-based model with segmentation and generation processes to address the error cascading problem. Su et al. (2021) proposed PlanGen, which applied a structured-aware reinforcement learning to cope with the cascading of errors. PlanGen uses the BLEU score between a gold sequence and a generated sequence as a reward to train the content planner so that PlanGen can directly train the content planner from the final output of the planning-based model. DYPLOC (Hua et al., 2021) uses multiple plan candidates to generate sentences in a text generator with content item conditioning based on the scores of the plan scoring network.

We employed a unified framework that enables both training time and inference time to cope with error cascading. In addition, our CoPlan addresses factual accuracy and the description order-based output candidate quality estimator for both time series irrelevant scenarios and time series critical scenarios to select the best plans.

3 Method

Radiology report generation is a task of generating reports comprising a sequence of sentences $Y = \{Y^1, ...Y^n\}$, where Y^i represents a sequence of words $Y^i = \{y^i_1, y^i_2, ...y^i_m\}$ from a set of images $X = \{x_k\}^M_{k=1}$. We annotated a set of finding labels $F = \{f_1, f_2, ...f_T\}$ for each set of images X. Further, we also annotated a subset of the described finding labels $F^i = \{f^1, ...f^k\}$ for each sentence Y^i. We define the plan F_{plan} as a list of subsets of finding labels $F_{plan} = \{F^1, ...F^i, ...F^k\}$.

3.1 Text Generation system with Planning

Unlike in an end-to-end system $Y = \mathrm{P}_{e2e}(X)$, our pipeline radiology report generation system comprises three stages; image classifier (IC), content planner (CP), and the text generator (TG). Image classifier $F = \mathrm{P}_{cv}(X)$ is a multi-class multi-label image classifier pretrained on ImageNet (Deng et al., 2009) to distinguish the finding labels F found in the image.

The content planner $F_{plan} = \mathrm{P}_{cp}(F)$ generates a plan F_{plan}, which represents the content and description order of the reports. We use an LSTM encoder-decoder with attention (Bahdanau et al., 2015) as the content planner. We employ three simple constraints (Shen et al., 2020) during the inference phase of the decoder of content planner:

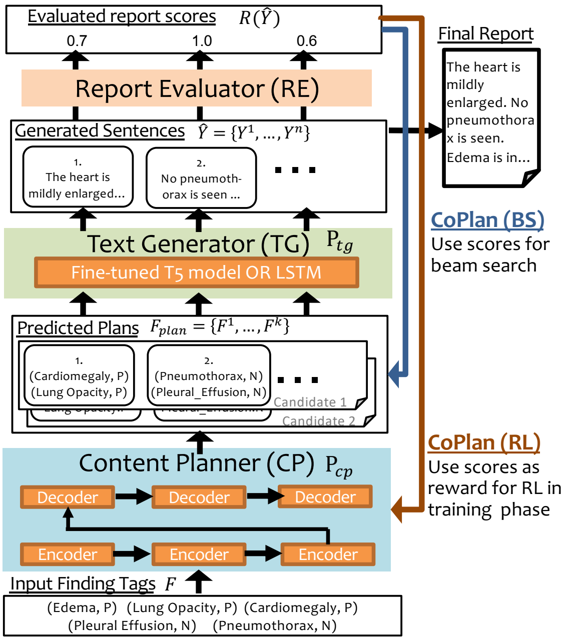

Evaluated report scores $R(\hat{Y})$
0.7
1.0
0.6
Report Evaluator (RE)
Generated Sentences $\hat{Y} = \{Y^1, ..., Y^n\}$
1. The heart is mildly enlarged...
2. No pneumoth-orax is seen ...
Text Generator (TG) P_{tg}
Fine-tuned T5 model OR LSTM
Predicted Plans $F_{plan} = \{F^1, ..., F^k\}$
1. (Cardiomegaly, P) (Lung Opacity, P)
2. (Pneumothorax, N) (Pleural_Effusion, N)
Candidate 1
Candidate 2
Content Planner (CP) P_{cp}
Decoder
Decoder
Decoder
Encoder
Encoder
Encoder
Input Finding Tags F
(Edema, P) (Lung Opacity, P) (Cardiomegaly, P) (Pleural Effusion, N) (Pneumothorax, N)
Final Report
The heart is mildly enlarged. No pneumothorax is seen. Edema is in...
CoPlan (BS)
Use scores for beam search
CoPlan (RL)
Use scores as reward for RL in training phase

Figure 3: An overview of the data-to-text generator with our proposed Coordinated Planning-based method (CoPlan). We leverage the estimated report scores for the training and inference of the content planner.

1. Segments in a plan must not be empty.
2. The same finding label cannot be realized more than once.
3. The generation of a plan must not be completed until all input finding labels have been realized.

Constraints 2 and 3 contribute to the significant reduction of the repetition problem and the missing information problem because all finding labels in the input are guaranteed to be generated only once in the generated plan.

The text generator $Y^i = \mathrm{P}_{tg}(F^i)$ generates each sentence Y^i from the predicted plan F^i sentence-by-sentence. We use T5 (Raffel et al., 2020) as the text generator. We treat the combination of the content planner and the text generator as the data-to-text (D2T) generation system $Y = \mathrm{P}_{d2t}(F)$.

3.2 Coordinated Planning-based text generation (CoPlan).

We propose Coordinated Planning-based text generation (CoPlan), which trains and conducts inference of the content planner and the text generator coordinately to address the error cascading problems. CoPlan comprises two methods: (1) reinforcement learning during the training phase and (2) coordinated inference with the beam search. Figure 3 shows an overview of CoPlan.

CoPlan in Reinforcement Learning (CoPlan (RL)). The error cascading problems are attributed to the independent training of the content planner and the text generator. The training of the content planner does not consider the text generator's final output; therefore, the content planner can generate inappropriate plans that cause errors in the generated reports. Further, there are several plan candidates for one correct report. Reinforcement learning (RL) is appropriate for these characteristics of the plan generation. Liu et al. (2019a) applied RL with clinically coherent rewards to train the text generator.

We introduce CoPlan in Reinforcement Learning (CoPlan (RL)), which uses an estimated quality of the generated reports as a reward for the RL of the content planner. CoPlan (RL) leverages the estimated quality of the final output of the system, such as factual accuracy, to train the content planner, the first stage of the planning-based modules. Therefore CoPlan (RL) can train the content planner to alleviate the cascading of errors.

The procedure of CoPlan (RL) is as follows. First, the text generator generates the report based on the plan predicted in the content planner. Second, a report evaluator (RE) calculates the quality of the generated reports and then uses it as a reward for the RL of the content planner. We adopt SCST (Rennie et al., 2017) to approximate this loss as:

$$L^{all} = \lambda_{rl} L^{rl} + (1 - \lambda_{rl}) L^{xent} \quad (1)$$

$$\nabla_\theta L_\theta^{rl} \approx - \nabla_\theta \log P_\theta(\hat{Y}^s)(\mathrm{R}(\hat{Y}^s) - \mathrm{R}(\hat{Y}^g)) \quad (2)$$

where L^{xent} indicates a cross-entropy loss, $\hat{Y}^s$ represents a sequence generated by a Monte Carlo sampling, $\hat{Y}^g$ is a sequence greedily generated, $\mathrm{R}(\hat{Y})$ represents the reward regarding the generated report $\hat{Y}$, and λ_{rl} is a hyperparameter.

We employ a report evaluator (RE) to quantify the quality of generated reports and use estimated scores as the reward $\mathrm{R}(\hat{Y})$. In this study, we use a reconstructor $\mathrm{REC}(\hat{Y})$ as the report evaluator. The reconstructor predicts the appropriate finding labels or description order from the generated reports in reverse, so it allows the report evaluator to quantify the clinical correctness.

We use two types of reconstructors to estimate the quality of the report: the factual accuracy-estimation reconstructor $\mathrm{REC}_{\mathrm{fact}}$ and the description order consistency-estimation reconstructor $\mathrm{REC}_{\mathrm{ord}}$. For the factual accuracy-estimation reconstructor $\mathrm{REC}_{\mathrm{fact}}$, we use fine-tuned ELECTRA (Clark et al., 2019) to predict finding labels from the reports, and an F-score of the predicted finding labels against the input finding labels is used as a report score. For the description order consistency-estimation reconstructor $\mathrm{REC}_{\mathrm{ord}}$, we use fine-tuned T5 to predict the description order of finding labels. The Damerau-Levenshtein Distance (Brill and Moore, 2000) between the sequence of input finding labels and the predicted labels are treated as a report score.

To stabilize the training of RL, we append the ROUGE scores to the reward to avoid the sparsity of the reward.

The overall reward $\mathrm{R}(\hat{\mathrm{Y}})$ regarding generated report $\hat{Y}$is formulated as follows:

$$\begin{aligned} \mathrm{R}(\hat{\mathrm{Y}}) = \quad & \lambda_{rouge} \mathrm{ROUGE}(Y, \hat{Y}) \\ & + (1 - \lambda_{rouge}) \mathrm{REC}(\hat{Y}) \end{aligned} \quad (3)$$

CoPlan in Beam Search (CoPlan (BS)). In addition to the RL, we introduce CoPlan (BS) in which the content planner decodes plans using the beam search in the inference phase in a coordinated manner. The output correctness is crucial for the practical use of medical systems, and thus, the system should avoid the risk of missing or incorrect descriptions. CoPlan (BS) aims to detect the errors in the outputs and correct them by modifying the plan used to generate them.

The content planner with CoPlan (BS) predicts the plan $\hat{F}_{plan} = \{\hat{f}_0, ..., \hat{f}_T\}$ in accordance with the factual accuracy or the consistent description order of the generated report $\hat{Y}$. The scores of report evaluator are added to the scoring function of the beam search. The recursive algorithm of the beam search is formulated as:

$$\begin{aligned} \hat{f}_0 &= < \mathrm{BOP} > \\ \hat{f}_t &= \underset{F'_{plan} \subseteq B}{\mathrm{argmax}} \log p_\theta(F'_{plan} | F) + \lambda_{re} \mathrm{RE}(\hat{Y}) \end{aligned} \quad (4)$$

where $\mathrm{RE}(\hat{Y})$ represents the scores of the report evaluator for $\hat{Y}$ generated by $\mathrm{P}_{tg}(\hat{F}_{plan})$ during the decoding step of beam search. λ_{re} is a hyperparameter. $\hat{f}_t$ denotes the predicted finding labels in time step t, and B indicates the candidate plans in the search space.

		Number of	Avg.	Avg.
Dataset	Split	Reports	Labels	Length
JLiverCT	Training	882	11.3	54.6
	Validation	127	12.3	63.7
	Test	74	12.2	62.4
MIMIC-CXR	Training	118,794	5.12	66.5
	Validation	1,196	5.82	56.6
	Test	2,347	6.45	64.0

Table 1: Statistics of the datasets.

$\mathrm{REC}_{\mathrm{fact}}$ and $\mathrm{REC}_{\mathrm{ord}}$ in CoPlan (RL) are also used as the report evaluator $\mathrm{RE}(\hat{Y})$ of CoPlan (BS).

4 Experiments

4.1 Datasets.

We used two datasets with different modalities and languages: JLiverCT for the time series-critical scenario and MIMIC-CXR for evaluating the time series irrelevant scenario. Table 1 presents the basic statistical features of the datasets. The details of these datasets and ethical policies are included in the Appendix for reproducibility.

The JLiverCT dataset. For the JLiverCT dataset, we collected radiology reports of liver lesions from a hospital and extracted 1,083 reports. All extracted reports had at least one description of findings regarding liver lesions and at least two descriptions describing a time series.

The JLiverCT dataset contains pairs of input sets for finding labels and target radiology reports written in Japanese. Following LI-RADS (Chernyak et al., 2018), we defined 65 types of finding labels and seven time series. We define the finding labels as a combination of time series, findings, and lesion conditions. For example, the presence of "ring enhancement" in the arterial phase is indicated as (Arterial, Ring_Enhancement, P), and the weak enhancement in the delayed phase is indicated as (Delayed, Enhancement, Weak). The time series represents a chronological order of scan timing: the first scan timing is "arterial phase," followed by the "early phase," "equilibrium phase," "delayed phase," and so forth. The status of the lesion indicates the degree of findings, such as "weak" or "strong," in addition to "positive" or "negative," which contributes to the estimated extent of the disease.

Annotators with sufficient knowledge of radiology reporting have manually annotated the finding labels in the reports. We focused only on the findings in the reports, so sentences unrelated to any finding labels were omitted from reports because of privacy concerns.

The MIMIC-CXR dataset. The MIMIC-CXR dataset includes chest X-ray images and the corresponding radiology reports written in English. We used the 14 categories of finding types defined in the CheXpert Labeler (Irvin et al., 2019). The original MIMIC-CXR dataset does not contain plan labels, so we have annotated the MIMIC-CXR dataset using the CheXpert labeler to obtain plans of the reports following the order that appeared in the report [1]. The finding label f_t is defined as a combination of finding type and polarity. Four polarity types are defined; abnormalities (indicated as P), normalities (indicated as N), uncertain findings (indicated as U), and no mentioned findings (indicated as X). For example, the label "(Pleural_Effusion, P)" is annotated to the report if the finding suggests pleural effusion.

Doctors are required to write concise and informative radiology reports, and they reflect their intention to write the reports by selecting the critical finding to be described in the reports or otherwise. In a few cases, doctors intentionally wrote normalities in the report to emphasize the absence of the finding, and in other cases, doctors intentionally omit the description regarding normalities. The former case is labeled as "negative" findings, while the latter is labeled as "no mention."

4.2 Models

Details of models, hyperparameter searches, training procedures, and the accuracy of reconstructors trained in advance are described in the Appendix C.

Models for the JLiverCT Dataset. We use LSTM as the content planner, T5 [2] as the text generator and the reconstructor $\mathrm{REC}_{\mathrm{ord}}$, and ELECTRA [3] as the reconstructor $\mathrm{REC}_{\mathrm{fact}}$.

Models for the MIMIC-CXR Dataset. The image classification model (IC) for the MIMIC-CXR dataset is a four-class multi-label classification task which diagnoses four polarity types for each finding type from images. We trained the 4-class IC with the annotated 4-class MIMIC-CXR dataset. The 4-class IC predicts a set of probabilities of all four types of polarity (abnormalities, normalities,

[1] Different to previous studies that used finding label annotations per report, we re-annotate the finding labels to obtain label annotations per each sentence with modified version the of CheXpert Labeler.

[2] megagonlabs/t5-base-japanese-web

[3] Cinnamon/electra-small-japanese-discriminator

uncertain, and no mention) for each finding label; it passes three types of labels other than the no mention label to the D2T module. We use EfficientNet-B4 (Tan and Le, 2019) in the multi-class multi-label classifier pretrained on ImageNet.

We use LSTM as the content planner, T5-base [4] as the text generator, and ELECTRA [5] as the reconstructor $\mathrm{REC}_{\mathrm{fact}}$. Only $\mathrm{REC}_{\mathrm{fact}}$ are used as the reconstructor of CoPlan because reports in the MIMIC-CXR dataset are time series irrelevant.

4.3 Definition of the Evaluation Metrics.

In addition to the NLG metrics, such as BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2019), and ROUGE (Lin, 2004), we deploy a clinical factual accuracy metric in Miura et al. (2021) [6] and Content Ordering (CO) metric (Wiseman et al., 2017) that quantifies the consistency of the description order of the reports. The description order not only indicates a chronological order but also relates to the clinical importance where the important findings of a report are likely to be written in the earlier parts of the report. CO metrics are commonly used to quantify the correctness of description order in the data-to-text research area, so we evaluated the generated reports with CO. The finding labels in the reports are extracted with the orders of the corresponding descriptions, and subsequently, CO is calculated as the normalized Damerau-Levenshtein Distance between the extracted labels of the predicted and the gold reports. CO can estimate the description order accurately based on the content of the reports rather than the surface-based metric in Kurisinkel et al. (2021). Details of these metrics, testing, and labeler are described in the Appendix C.

5 Experimental Results

We conducted two types of experiments: the D2T experiment for evaluating the D2T module and the E2E experiment for evaluating the entire system.

5.1 Effect of Planning and CoPlan

Result on the JLiverCT Dataset. We conducted a D2T experiment using the JLiverCT dataset to evaluate the effect of the planning and CoPlan on the *time-series critical scenario*. We applied two types of reconstructors for CoPlan: $\mathrm{REC}_{\mathrm{fact}}$ (indicated as $\mathrm{CoPlan}_{\mathrm{fact}}$) and $\mathrm{REC}_{\mathrm{ord}}$ (indicated as $\mathrm{CoPlan}_{\mathrm{ord}}$). We prepared three baseline models to calibrate our results: template-based generation (**Template**), nearest-neighbor search method (**1-NN**) (Boag et al., 2020), and T5 model in which the labels are fed in chronological order (**T5-base**).

An automatic evaluation result of Table 2 indicates that $\mathrm{CoPlan}_{\mathrm{ord}}$ achieved the best BLEU4 and CO scores among all and the best factual accuracy among all neural-based models. $\mathrm{CoPlan}_{\mathrm{ord}}$ predicts better plans than $\mathrm{CoPlan}_{\mathrm{fact}}$; this results in the improvement in the factual accuracy as well as content order scores. The template was the most factually accurate, but the corresponding reports were inappropriate because of unnecessary redundancy. Additionally, the average length was longer than text generation models. According to the radiologists consulted, they focus significantly on concise and consistent description order reports in addition to the factual accuracy; therefore, neural-based generation models are preferred. On the contrary, reports generated by T5 tend to be short and omit important descriptions. Because of the imbalanced nature of the JLiverCT dataset, a plain T5 model causes omission problems, but the planning models effectively reduce omissions.

Comparison of Report Evaluator Ablation study of Table 2 shows a comparison of report evaluator type between $\mathrm{REC}_{\mathrm{fact}}$ and $\mathrm{REC}_{\mathrm{ord}}$. $\mathrm{CoPlan(RL)}_{\mathrm{ord}}$ contributed to the improvement of both factual accuracy and content ordering scores, while $\mathrm{CoPlan(RL)}_{\mathrm{fact}}$ slightly improved factual accuracy. For the time series critical scenario, generated plans with appropriate ordering tend to be similar to the gold plans in training data, so this results in improving the factual accuracy of the text generator.

From Table 2, we assumed that RL with factual reward $\mathrm{CoPlan(RL)}_{\mathrm{fact}}$ has no effect on the factual accuracy. To investigate the effect of RL with factual reward, we further compared the plain planning-based model with $\mathrm{CoPlan(RL)}_{\mathrm{fact}}$ without the constraints mentioned in Sec 3.1. Without constraints, $\mathrm{CoPlan(RL)}_{\mathrm{fact}}$ improved the factual accuracy by 1.3 pt compared to Planning. Therefore $\mathrm{CoPlan(RL)}_{\mathrm{fact}}$ clearly improves the accuracy of reports, but the constraints concealed the effect.

Result on the MIMIC-CXR Dataset. Additionally, we conducted a D2T experiment using the

[4] t5-base

[5] kamalkraj/bioelectra-base-discriminator-pubmed

[6] These scores have some gaps from the scores reported on previous works because we conducted report-level evaluation to follow the results of Miura et al. (2021).

JLiverCT (Time Series Critical Report)							
	CoPlan Type	Reconstructor Score	ROUGE-L	BLEU4	Accu.	CO	Avg.Len
Baseline Models							
Template	-	-	37.1	36.3	**99.3**	36.8	98.1
1-NN	-	-	45.9	32.2	72.1	40.9	50.1
T5-base	-	-	57.3	47.2	78.3	49.7	44.9
Ablation Studies							
Planing	-	-	59.1	48.0	82.1	52.1	48.9
$\mathrm{CoPlan(RL)_{fact}}$	RL	$\mathrm{REC_{fact}}$	57.9	47.4	81.9	49.2	50.8
$\mathrm{CoPlan(BS)_{fact}}$	BS	$\mathrm{REC_{fact}}$	58.4	48.1	82.7	46.9	54.2
$\mathrm{CoPlan(RL)_{ord}}$	RL	$\mathrm{REC_{ord}}$	**60.4**	48.1	83.0	52.7	55.6
$\mathrm{CoPlan(BS)_{ord}}$	BS	$\mathrm{REC_{ord}}$	59.8	48.6	83.1	53.2	58.5
Ablation Studies (without Constraints in Sec. 3.1)							
Planing	-	-	49.8	39.3	70.8	43.8	44.7
$\mathrm{CoPlan(RL)_{fact}}$	RL	$\mathrm{REC_{fact}}$	51.1	39.1	72.1	44.3	40.2
Proposed approaches							
$\mathrm{CoPlan_{ord}}$	RL+BS	$\mathrm{REC_{ord}}$	*60.3	***48.9**	*83.8	***54.8**	56.1
$\mathrm{CoPlan_{fact}}$	RL+BS	$\mathrm{REC_{fact}}$	59.2	46.6	82.8	49.8	53.1
MIMIC-CXR (Time Series Irrelevant Report)							
	CoPlan Type	Reconstructor Score	ROUGE-L	BLEU4	Accu.	CO	Avg.Len
Baseline model							
T5-base	-	-	**12.4**	**17.3**	92.2	34.3	45.3
Ablation studies							
Planning	-	-	11.9	16.4	94.3	40.7	45.9
$\mathrm{CoPlan(RL)_{fact}}$	RL	$\mathrm{REC_{fact}}$	11.9	16.5	95.1	41.1	53.2
$\mathrm{CoPlan(BS)_{fact}}$	BS	$\mathrm{REC_{fact}}$	12.3	17.0	95.9	45.8	50.9
Proposed approach							
$\mathrm{CoPlan_{fact}}$	RL+BS	$\mathrm{REC_{fact}}$	12.2	16.8	***97.6**	***47.7**	48.3

Table 2: Automatic evaluation results of the D2T experiment. Accu. and CO indicates the factual accuracy and consistency of the description order, and Avg.Len indicates the average length of generated reports. **Bold** are the best results, and scores with * are statistically significant compared to the baseline T5-base ($p < 0.05$).

MIMIC-CXR dataset to evaluate the effect of the planning and CoPlan with $\mathrm{REC_{fact}}$ on the *time-series irrelevant scenario*. We compared our CoPlan with the T5-base and a plain planning model without CoPlan. We applied only reconstructor $\mathrm{REC_{fact}}$ for CoPlan (indicated as $\mathrm{CoPlan_{fact}}$) because the factual accuracy is the most critical for the time series irrelevant scenario.

Table 2 shows the results of the D2T experiment. Both $\mathrm{CoPlan(RL)_{fact}}$ and $\mathrm{CoPlan(BS)_{fact}}$ improve factual accuracy; $\mathrm{CoPlan_{fact}}$ further improves factual accuracy. However, the surface-based metrics (BLEU4, ROUGE) are slightly decreased. The differences in BLEU and ROUGE are because the reports of CoPlan are redundant, as shown in Avg.Len in Table 2.

These results on the JLiverCT and MIMIC-CXR indicate that our CoPlan improves the quality of generated reports for both time series critical and irrelevant scenarios, provided that a report evaluator is selected correctly.

5.2 Comparison with Previous Studies

We conducted an E2E experiment on the MIMIC-CXR dataset to evaluate the entire system. We compare our CoPlan with four previous studies: **CoAtt** (Jing et al., 2018), which comprises of hierarchical LSTM with auxiliary tag prediction task, **R2Gen** (Chen et al., 2020), which uses memory-driven transformer, **IFCC** (Miura et al., 2021), which applies RL with NLI-based rewards, **R2GenCMN** (Chen et al., 2021), which deploys a cross-modal memory network to enhance encoder-decoder model, and **R2GenRL** (Qin and Song, 2022), which applied RL with NLG metrics.

Table 3 shows a result of the E2E experiment. In surface-based metrics, such as BLEU scores, our proposed system has a slightly lower score than R2Gen and IFCC. However, in clinical-based metrics, our proposed system improves the scores of factual accuracy and CO. The results show that our proposed model can generate reports with a more correct and consistent description order compared to end-to-end systems.

5.3 Human Evaluation

We conducted a human evaluation to validate the effect of CoPlan. We used three human evaluation metrics for the JLiverCT dataset: correctness, fluency, and content order metrics. Only correctness and fluency are used for the MIMIC-CXR dataset because of the time series irrelevant scenario. Cor-

E2E Model		Surface Metrics		Clinical Accuracy(Miura et al., 2021)				Label Type Accu.			CO
IC	D2T	BLEU4	BERTScore	Precision	Recall	Micro-F	Accuracy	Abn.	Nor.	Unc.	
Gold IC	CoPlan	12.2	60.6	97.8	96.9	97.6	97.7	97.1	98.2	92.3	47.7
4-class IC	T5-base (Baseline)	7.6	50.2	42.3	59.7	49.5	65.2	48.9	53.3	16.2	15.2
2-class IC	CoPlan	7.4	48.9	35.1	44.2	39.2	58.8	**57.6**	19.2	7.9	17.2
4-class IC	CoPlan (Proposed)	8.0	**57.1**	***68.1**	51.4	*__58.6__	*__77.1__	55.4	**67.2**	17.2	***27.9**
CoAtt (Jing et al., 2018)		7.8	51.7	46.8	35.1	40.1	74.2	33.9	54.7	12.4	21.1
R2Gen (Chen et al., 2020)		8.6	50.8	41.2	29.8	34.6	73.9	35.7	50.0	16.5	19.1
IFCC (Miura et al., 2021)		**11.4**	56.9	50.3	**65.1**	56.7	**77.1**	56.6	55.4	16.7	18.5
R2GenCMN (Chen et al., 2021)		9.6	53.3	55.4	44.8	49.5	72.8	44.1	51.9	10.0	20.5
R2GenRL (Qin and Song, 2022)		10.1	54.9	56.1	46.2	50.7	74.0	45.3	52.2	10.7	19.8

Table 3: Automatic evaluation results of the E2E experiment on MIMIC-CXR dataset. Precision, Recall, Micro-F, and Accuracy represent the clinical accuracy scores output by the clinical CheXBERT labeler. CO represents the clinical content ordering score. Abn., Nor., and Unc. indicate the Micro-F scores of abnormalities, normalities and uncertain finding labels, respectively. The scores with * are statistically significant compared with the baseline model ($p < 0.01$).

	Correctness	Fluency	CO
JLiverCT Dataset (Time Series Critical Report)			
T5-Japanese-base	86.5	4.48	61.6
CoPlan	**89.8**	**4.56**	**68.9**
MIMIC-CXR Dataset (Time Series Irrelevant Report)			
4-class IC + T5-base	58.4	**4.85**	-
4-class IC + CoPlan	**63.2**	4.79	-

Table 4: Results of a human evaluation results on the D2T experiment with the JLiverCT (upper) and the MIMIC-CXR (lower). All Krippendorff's $\alpha \geq 0.61$, with specific values in Appendix C.

rectness measures how well a report describes its clinical information. We define the correctness of reports as an F-score between the finding labels observed in a generated report and the labels contained in the corresponding gold report. The fluency score evaluates the naturalness of the generated reports with a 5-point Likert scale. Annotators extract the positions of descriptions regarding any finding labels in the gold report and the generated report; then, these positions are used to calculate the normalized Damerau-Levenshtein Distance to obtain the content order score. Two experts for the JLiverCT and six experts for the MIMIC-CXR dataset who are knowledgeable in radiology reports measured 100 randomly selected reports, as in previous research (Zhang et al., 2020).

Table 4 shows a human evaluation result in the JLiverCT and the MIMIC-CXR dataset. Our proposed CoPlan is effective on both the correctness and content order of the generated reports; however, the fluency is slightly decreased in the MIMIC-CXR dataset. The redundancy of the generated reports caused this drop of the fluency.

6 Discussion

6.1 Qualitative Results

The middle section of Table 5 presents examples of the generated reports with T5, Planning without CoPlan, and CoPlan of the JLiverCT dataset in the D2T experiment. In the reports generated using T5, several descriptions of the input finding label are omitted. The report by Planning without CoPlan has no omissions or missing findings, but the repetition resulted from the poor plan. The two sentences regarding the findings in the same phase "arterial phase" should be combined to one sentence to generate a concise and informative report. However, the reports generated by CoPlan are written in a consistent order without any omission. This shows CoPlan can generate reports with a description order consistent with the gold report.

6.2 Importance of Normality and Uncertain Labels.

We further analyzed the clinical accuracy scores for each type of finding label to observe the differences in the modality of the finding labels. In addition to the 4-class IC in Sec 4.1, we employed a 2-class IC trained by the 2-class MIMIC-CXR dataset. The 2-class MIMIC-CXR dataset was annotated with VisualCheXBERT (Jain et al., 2021b), and two polarity labels were annotated: positive or negative findings. The 2-class IC predicts a set of probabilities of two types of labels (positive and negative) for each finding label; it passes only the positive labels as abnormalities to the D2T module.

Table 6 shows a comparison of the 4-class IC and the 2-class IC. A large discrepancy between 4-class IC and 2-class IC indicates a difference between the findings shown in the radiology images

Input Labels of the D2T module. (Arterial, Enhancement, P), (Arterial, Enhancement, Strong), (Portal, Enhancement, P), (Portal, Enhancement, Persistent), (Delayed, Enhancement, P), (Delayed, Enhancement, Weak), (Delayed, Enhancement, Persistent), (No_Phase, Lesion, Hypervascular_Type)
Generated Report in Japanese-T5-base (Baseline) In S2, 10 cm lesion with strong enhancement is observed in the arterial phase. Persistent enhancement is observed in the portal phase. Strong enhancement is observed in the the delayed phase. It is hypervascular lesion.
Generated Report in Gold-IC + Planning (CoPlan not applied) In S2, 10 cm lesion with enhancement is observed in the arterial phase. Strong enhancement is observed in the arterial phase. Persistent enhancement is observed in the portal phase. Weak persistent enhancement staining in the delayed phase. It is a hypervascular lesion.
Generated Report in Gold-IC + CoPlan (Proposed) In S2, 10 cm lesion with strong enhancement is observed in the arterial phase. There is persistent enhancement staining in the portal phase. Weak persistent enhancement staining in the delayed phase. It is a hypervascular lesion.

Table 5: Examples of generated reports of the JLiverCT in the D2T experiment (Translated).

	Micro-F	Abn.	Nor.	Unc.	Not.
2-class IC	80.2	62.3	-	-	81.9
4-class IC	67.6	56.2	64.6	16.8	75.7

Table 6: Results of image classification evaluation on MIMIC-CXR dataset. Abn., Nor., Unc., and Not. indicate the Micro-F scores of abnormalities, normalities, uncertain, and not mentioned finding labels.

and those described in the reports. For a concise and informative report, radiologists intentionally omit some apparent findings and obscure descriptions (Jain et al., 2021b), and thus, the findings described in the reports deviate from the findings in the images, particularly for the 4-class IC.

The right section of Table 3 shows the factual accuracy of the reports generated by the label type. Gold IC + CoPlan indicates the upper bound of the performance of the D2T module when the gold classification results are provided. The results of the 4-class IC + CoPlan and 2-class IC + CoPlan are significantly lower than those of the Gold IC + CoPlan; this is because it is difficult for IC to distinguish normalities and uncertain labels significantly affected by the intentions of radiologists. The 2-class IC cannot predict negative and uncertain labels, and therefore, the 2-class IC + CoPlan merely generates the description of normalities and uncertain findings. Regarding the adequacy of abnormalities, the 4-class IC + CoPlan is lower than that of the 2-class IC + CoPlan because the 4-class IC is severely affected by the presence of normalities and uncertain findings.

This result indicates that the all fully-automated radiology report generation systems have a limitation to generate descriptions about normalities and uncertain finding labels without the intentions of doctors. The radiology report generation systems must comply with the intentions of doctors to correctly generate descriptions before applying the radiology report generation systems in practice.

7 Conclusion

We proposed a planning-based neural radiology report generation method for generating reports with the consistent description order on top of the factual accuracy of the content.The results of the evaluations in both time series critical and time series irrelevant datasets revealed that our proposed CoPlan improved both the factual accuracy and consistency of the description order of the generated reports. However, as shown in Sec. 6.2, all radiology report systems have a limitation to generate descriptions regarding normalities without doctors' intentions. In the future, we will combine our system with a human-in-the-loop approach that can reflect doctors' intentions to co-create high-quality reports in a short time.

A Limitations

We recognize that this system currently targets *support the workflow of radiologists*, not substituting the role of a radiologist in the entire workflow. From the discussion of Sec. 6.2, the report generation systems without any intervention from doctors have an obstacle to generating reports in which the intention of doctors is adequately reflected. A human-in-the-loop system that enables the workflow composed of suggested generated candidate reports, corrects predicted plans by radiologists, subsequently completes reports, reflecting the intention of doctors. Our planning-based radiology report generation system can easily build a human-in-the-loop system that can reflect doctors' intentions because it uses discrete representations for the plans; this is a great advantage of our approach compared to the existing systems. Radiologists can check and correct the result of the image classifier or the content planner, and this strategy excessively reduces the risk that the system errors could threaten the life of patients while contributing to the reduction of the radiologists' workload.

B Ethics Statement

Both the JLiverCT dataset and the MIMIC-CXR dataset were de-identified to respect patients' privacy. We use the MIMIC-CXR dataset under the license of PhysioNet Credentialed Health Data License 1.5.0 [7]. On the distributed MIMIC-CXR dataset, all Protected Health Information (PHI) was removed to satisfy the US Health Insurance Portability, and Accountability Act of 1996 (HIPAA) Safe Harbor requirements (Johnson et al., 2019). Likewise, on our originally collected JLiverCT dataset, all personal information in the reports was removed to respect patients' privacy. We extracted descriptions referring only to findings, and all other descriptions including medical examination numbers and names of the patients are omitted. All radiographs and radiology reports used to construct the JLiverCT dataset were collected under the agreement of patients or agents of patients, and the JLiverCT dataset and this research have been approved by the Institutional Review Board of the hospital and our institution.

[7] https://physionet.org/content/mimic-cxr/view-license/2.0.0/

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *The 2015 International Conference on Learning Representation.*

William Boag, Tzu-Ming Harry Hsu, Matthew McDermott, Gabriela Berner, Emily Alesentzer, and Peter Szolovits. 2020. Baselines for chest x-ray report generation. In *Machine Learning for Health Workshop*, pages 126–140. PMLR.

Eric Brill and Robert C Moore. 2000. An improved error model for noisy channel spelling correction. In *Proceedings of the 38th annual meeting of the association for computational linguistics*, pages 286–293.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5904–5914, Online. Association for Computational Linguistics.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449.

Victoria Chernyak, Kathryn J Fowler, Aya Kamaya, Ania Z Kielar, Khaled M Elsayes, Mustafa R Bashir, Yuko Kono, Richard K Do, Donald G Mitchell, Amit G Singal, et al. 2018. Liver imaging reporting and data system (li-rads) version 2018: imaging of hepatocellular carcinoma in at-risk patients. *Radiology*, 289(3):816.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2019. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *International Conference on Learning Representations.*

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277.

Jia Deng, Wei Dong, R 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.*

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392.

European Society of Radiology (ESR). 2011. Good practice for radiological reporting. guidelines from the european society of radiology (esr). *Insights into Imaging*, 2:93–96.

Thiago Castro Ferreira, Chris van der Lee, Emiel Van Miltenburg, and Emiel Krahmer. 2019. Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 552–562.

Xinyu Hua, Ashwin Sreevatsa, and Lu Wang. 2021. DYPLOC: Dynamic Planning of Content Using Mixed Language Models for Text Generation. In *Proceedings of the 27th Annual Meeting of the Association for Natural Language Processing*.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Thirty-Third AAAI Conference on Artificial Intelligence*.

Saahil Jain, Akshay Smit, Andrew Y Ng, and Pranav Rajpurkar. 2021a. Effect of radiology report labeler quality on deep learning models for chest x-ray interpretation. *arXiv preprint arXiv:2104.00793*.

Saahil Jain, Akshay Smit, Steven QH Truong, Chanh DT Nguyen, Minh-Thanh Huynh, Mudit Jain, Victoria A Young, Andrew Y Ng, Matthew P Lungren, and Pranav Rajpurkar. 2021b. Visualchexbert: addressing the discrepancy between radiology report labels and image labels. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 105–115.

Baoyu Jing, Pengtao Xie, and Eric Xing. 2018. On the Automatic Generation of Medical Imaging Reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. MIMIC-CXR: A large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Litton J Kurisinkel, Aiti Aw, and Nancy F Chen. 2021. Coherent and concise radiology report generation via context specific image representations and orthogonal sentence states. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 246–254.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of Workshop on Text Summarization Branches Out*.

Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. 2019a. Clinically Accurate Chest X-Ray Report Generation. In *Machine Learning for Healthcare Conference 2019*.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2019b. On the Variance of the Adaptive Learning Rate and Beyond. *arXiv preprint arXiv:1908.03265*.

Shuming Ma, Pengcheng Yang, Tianyu Liu, Peng Li, Jie Zhou, and Xu Sun. 2019. Key fact as pivot: A two-stage model for low resource table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2047–2057.

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. 2021. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5288–5304.

Maram Mahmoud A Monshi, Josiah Poon, and Vera Chung. 2020. Deep learning in generating radiology reports: A survey. *Artificial Intelligence in Medicine*, page 101878.

Amit Moryossef, Yoav Goldberg, and Ido Dagan. 2019. Step-by-Step: Separating Planning from Realization in Neural Data-to-Text Generation. 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)*.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? *Advances in neural information processing systems*, 32.

Hoang Nguyen, Dong Nie, Taivanbat Badamdorj, Yujie Liu, Yingying Zhu, Jason Truong, and Li Cheng. 2021. Automated generation of accurate & fluent medical X-ray reports. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3552–3569, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Toru Nishino, Ryota Ozaki, Yohei Momoki, Tomoki Taniguchi, Ryuji Kano, Norihisa Nakano, Yuki Tagawa, Motoki Taniguchi, Tomoko Ohkuma, and Keigo Nakamura. 2020. Reinforcement learning with imbalanced dataset for data-to-text medical report generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2223–2236.

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 on Association for Computational Linguistics*.

Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. 2018. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings*, 2018:188.

Han Qin and Yan Song. 2022. Reinforced cross-modal alignment for radiology report generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 448–458.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*.

Toshinori Sato, Taiichi Hashimoto, and Manabu Okumura. 2017. Implementation of a word segmentation dictionary called mecab-ipadic-NEologd and study on how to use it effectively for information retrieval. In *Proceedings of the 23th Annual Meeting of the Association for Natural Language Processing*.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

Xiaoyu Shen, Ernie Chang, Hui Su, Cheng Niu, and Dietrich Klakow. 2020. Neural data-to-text generation via jointly learning the segmentation and correspondence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7155–7165.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew Lungren. 2020. CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519.

Yixuan Su, David Vandyke, Sihui Wang, Yimai Fang, and Nigel Collier. 2021. Plan-then-generate: Controlled data-to-text generation via planning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 895–909.

Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*.

Chen Tang, Frank Guerin, Yucheng Li, and Chenghua Lin. 2022. Recent advances in neural text generation: A task-agnostic survey. *arXiv preprint arXiv:2203.03047*.

Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2017. Challenges in Data-to-Document Generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. 2020. When Radiology Report Generation Meets Knowledge Graph. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

C Appendix

C.1 Dataset and Preprocessing

JLiverCT Dataset. We constructed the JLiverCT dataset to train the data-to-text module of the radiology report generation system. We collected 1,083 reports that indicate the diagnosis of liver contrast CT from a hospital. In the preprocessing phase, we omitted the sentences that did not describe the CT images' findings to avoid violating patients' privacy. We annotated 65 types of finding labels, seven instances of time series, and lexicalized descriptions referring to the position and size of the nodules For training an LSTM-based text generation model, we used MeCab [8] and mecab-ipadic-NEologd (Sato et al., 2017), to tokenize the reports, and for training the T5-based model, we used SentencePiece-based tokenizer (Kudo and Richardson, 2018) trained on the Japanese Wikipedia dataset.

the MIMIC-CXR Dataset. We used the MIMIC-CXR Dataset [9], which contains pairs of chest X-ray radiographs and free-text radiology reports. In the preprocessing phase, we extracted the finding sections of the reports using the scripts [10] and split the reports into train, validation, and test data based on the split distributed in the MIMIC-CXR-JPG (Johnson et al., 2019) [11] dataset. In the training data, we truncated the sentences in the reports that were unrelated to any findings using the CheXpert Labeler and NegBio (Peng et al., 2018) parser to improve the stability of training the model. We omitted reports that did not mention any findings or had no finding sections from the training data.

MIMIC-CXR Dataset for Image Classifier. To train the image classification module (IC), we annotated the MIMIC-CXR dataset with two automated labelers: VisualCheXBERT (Jain et al., 2021b), CheXBERT (Smit et al., 2020), and CheXpert Labeler (Irvin et al., 2019). We annotate a 4-class image classification dataset with CheXBERT, which can annotate 14 categories of labels with four polarities: abnormalities, normalities, uncertain, and unseen. However, Jain et al. (2021a) reported that the accuracy of annotations with CheXpert Labeler is lower than that of a system including human expert annotations, in terms of normalities and uncertain labels. We annotated a 2-Class image classification dataset with VisualCheXBERT which can annotate 14 categories of labels with two polarities: abnormalities and other than abnormalities. Jain et al. (2021b) trained VisualCheXBERT with both report labels and image labels to annotate more accurate labels. VisualCheXBERT adopts the ZeroOne strategy, which maps the uncertain and unseen labels to positive (abnormalities) or negative labels. Therefore, VisualCheXBERT can annotate labels more accurately than CheXpert Labeler, but it cannot annotate normalities and uncertain labels.

	Hyperparameters
Max Epochs	8
Batch Size	4, **8**, 16
Learning Rate	1e-5, 3e-5, **1e-4**, 3e-4, 1e-3
Gradient Clipping	**1.0**, 2.0, ∞
Optimizer	Adam, **RAdam**, AdaFactor
Label Smoothing	0.0, **0.1**, 0.2

Table 7: The hyperparameters tested in tuning our image classifier. **Bold** values indicate the best hyper-parameter configuration.

the MIMIC-CXR Dataset for the Data-to-Text module. We annotated finding labels and plans to the MIMIC-CXR dataset with the CheXpert Labeler (Irvin et al., 2019) to train the data-to-text module. We define the plans F_{plan} as follows: $F_{plan} = \{F^1, F^2 ... F^k\}$,$F^i = \{f^1, ... f^j\}$ for each sentence Y^i. For example, we annotated the plan "(Lung_Opacity, P) <SEP> (Pleural_Effusion, N)"to the report "There is a new opacity in the left lobe. No pleural effusion."

However, the original CheXpert Labeler cannot extract the finding labels with their positions mentioned in the report. We modified the extraction process inside the CheXpert Labeler to include the described position of a lesion as written in the report and annotated the labels accordingly for each sentence in the report. We omitted annotated sentences with no finding labels because these descriptions cannot be generated from input images.

For calculating CO metrics, we utilized the CheXpert Labeler for MIMIC-CXR, and the original rule-based labeler was used for the JLiverCT to extract the finding labels with the corresponding descriptions' orders.

C.2 Training Details

Image Classifier (IC). All images were fed into a network of the size of 256×256 pixels. We defined the loss as the sum of the multi-class class-balanced cross-entropy loss (Cui et al., 2019) and used the

[8] https://taku910.github.io/mecab/

[9] https://physionet.org/content/mimic-cxr/2.0.0/

[10] https://github.com/MIT-LCP/mimic-cxr/

[11] https://physionet.org/content/mimic-cxr-jpg/2.0.0/

dataset	JLiverCT	MIMIC-CXR
Model Hyperparameters		
Dropout rate	0.1, **0.15**, 0.3	0.1, **0.15**, 0.3
Label embedding size	**16**, 32	**16**, 32
Hidden size	**32**, 64, 128	**32**, 64, 128
Beam search width	3	3
Training Hyperparameters		
Max Epochs	**50**	**20**
Batch size	8, **16**, 32	8, 16, **32**
Optimizer	**Adam**, SGD	**Adam**, SGD
Learning rate	1e-3, 2e-3, **1e-2**	1e-3, **2e-3**, 1e-2
Learning rate decay	0.95, **0.98**, 1.0	0.95, **0.98**, 1.0
λ_{rouge}	0.0, **0.05**, 0.1, 0.3	0.0, 0.05, **0.1**, 0.3
λ_{rl}	0.1, **0.2**, 0.5, 0.8	**0.1**, 0.2, 0.5, 0.8
λ_{re}	0.1, 0.2, **0.2**, 1.0	0.1, 0.2, **0.5**, 1.0
Gradient clipping	1.0, **2.0**, ∞	1.0, **2.0**, ∞

Table 8: The hyperparameters tested in tuning our content planner. **Bold** values indicate the best hyper-parameter configuration.

dataset	JLiverCT	MIMIC-CXR
Training Hyperparameters		
Max Epochs	20	5
Batch size	2, 3, **4**	2, 4, **6**
Optimizer	**AdaFactor** (Shazeer and Stern, 2018)	**AdaFactor**
Learning rate	1e-4, 3e-3, **1e-3**	**1e-4**, 3e-3, 1e-3
Learning rate decay	0.95, **0.98**, 1.0	**0.95**, 0.98, 1.0
Dropout	0.1, **0.15**, 0.3	0.1, **0.15**, 0.3
Gradient clipping	1.0, **2.0**, ∞	1.0, **2.0**, ∞
Accumulate Batches	**2**	**3**
Beam Width	**3**	**3**

Table 9: The hyperparameters tested in tuning our text generator. **Bold** values indicate the best hyper-parameter configuration.

RAdam (Liu et al., 2019b) optimizer with a learning rate of 1.0×10^{-4}. We applied label smoothing (Müller et al., 2019) with the hyperparameter $\alpha = 0.1$. Table 11 presents hyperparameters used to train the image classifier. We manually tuned all hyperparameters on the validation set of the MIMIC-CXR dataset, and the models with highest F-scores $\mathrm{REC}(\hat{\mathrm{Y}}^{\mathrm{g}})$ were selected as the best model. Table 10 presents F-scores for each finding label in the MIMIC-CXR dataset.

It is worth mentioning that specific studies include two or more diagnostic images (e.g., frontal and lateral images) in one report. First, the image classifier estimates the predicted scores of finding labels for each image in one study, and then, the average scores of the images are calculated to obtain the classification result for the entire study. To deal with the imbalanced nature of the MIMIC-CXR dataset, we optimized the threshold of the output probability scores of the classification model for each finding label. We evaluated the validation set with the threshold values between 0.0 to 1.0 in increments of 0.05 and then determined the threshold values which achieved the best F-scores as the threshold for the IC module. Table 11 presents hyperparameters used to train the image classifier. We manually tuned all hyperparameters on the validation set of the MIMIC-CXR dataset, and the model with highest F-score was selected as the best model.

Data-to-Text Module We used T5-small model provided by Huggingface [12] for the text generator of the MIMIC-CXR and T5-Japanese-base model[13] for the text generator of the JLiverCT dataset. Table 8 and Table 9 present hyperparameters used to train the content planner and the text generator. We manually tuned all hyperparameters on the validation set of the datasets, and the models with highest report evaluator scores $\mathrm{REC}(\hat{\mathrm{Y}}^{\mathrm{g}})$ were selected as the best model. The number of parameters of the data-to-text module was 220M for the JLiverCT dataset and 61M for the MIMIC-CXR dataset. We used an Intel Core i9-9900K CPU and NVIDIA GTX 2080 GPU for training, and the training time was approximately 12h for the JLiverCT dataset and 40h for the MIMIC-CXR dataset. The ROUGE-L score of our CoPlan on the validation set of the JLiverCT dataset is 60.6, and the ROUGE-L score of our CoPlan on the validation set of the JLiverCT dataset is 12.5, respectively.

Reconstructor. We used the pretrained Japanese BERT model [14] to train the factual accuracy reconstructor $\mathrm{REC}_{\mathrm{fact}}$ and Japanese T5 model [15] to train the description order reconstructor $\mathrm{REC}_{\mathrm{ord}}$ for the JLiverCT dataset. We split the training data contained in the data-to-text module into 4:1 ratio and used the greater part as training data and the smaller part as validation data for the reconstructor. We used binary cross-entropy loss to train the model and applied Class Balanced Loss (CBL) (Cui et al., 2019) with $\beta = 0.999$ to the BERT model $\mathrm{REC}_{\mathrm{fact}}$. The number of parameters of the reconstructor was 110M for $\mathrm{REC}_{\mathrm{fact}}$, and 247M for $\mathrm{REC}_{\mathrm{ord}}$. We fine-tuned the model with five epochs and conducted 5-fold cross-validation to determine the hyperparameters. The F-score on the validation dataset was 99.4 for $\mathrm{REC}_{\mathrm{fact}}$ and 98.1 for $\mathrm{REC}_{\mathrm{ord}}$. We used an Intel Core i9-9900K CPU and NVIDIA GTX 2080 GPU for training, and the

[12] https://huggingface.co/t5-small

[13] https://huggingface.co/sonoisa/t5-base-japanese

[14] https://github.com/cl-tohoku/bert-japanese

[15] https://github.com/megagonlabs/t5-japanese

Labels	Positive	Negative	Uncertain	No_Mention
Enlarged_Cardiomediastinum	1.48	19.7	27.3	53.7
Cardiomegaly	71.4	55.7	0.0	36.1
Lung_Opacity	61.1	50.0	0.0	54.4
Lung_Lesion	25.4	0.0	0.0	89.5
Edema	57.0	0.0	3.57	68.2
Consolidation	23.6	32.3	0.0	60.6
Pneumonia	30.0	56.3	20.6	69.6
Atelectasis	51.4	21.3	0.0	76.5
Pneumothorax	0.0	81.0	0.0	0.0
Pleural_Effusion	74.7	80.0	0.0	1.9
Pleural_Other	27.0	0.0	0.0	95.4
Fracture	18.7	10.0	0.0	81.9
Support_Devices	71.2	0.0	0.0	81.6
No_Finding	0.0	-	-	97.3
Overall F1-Score	56.2	64.6	16.8	75.7

Table 10: F-scores of the results of the 4-class image classifier (4-class IC) for each finding label.

Reconstructor Type	REC_{fact} (JLiverCT)	REC_{ord} (JLiverCT)	REC_{fact} (MIMIC-CXR)
Pretrained Model	cl-tohoku/bert-base-japanese	megagonlabs/ t5-base-japanese-web	google/electra-base-discriminator
Optimizer	AdamW	AdaFactor	AdamW
Learning rate of pretrained model layer	1e-5, **2e-5**, 1e-4	1e-5, 2e-5, **1e-4**	1e-5, **2e-5**, 1e-4
Learning rate of FC layer	1e-4, **2e-4**, 1e-3	**1e-4**, 2e-4, 1e-3	1e-4, **2e-4**, 1e-3
CBL β (Cui et al., 2019)	0, 0.99, **0.999**	-	0, 0.99, **0.999**
Warm up steps	0, **50**, 500	0, **50**, 500	0, 100, **1000**

Table 11: The hyperparameters tested in tuning our reconstructor. **Bold** values indicate the best hyper-parameter configuration.

training time was approximately two hours.

We used the pretrained ELECTRA-based model to train the reconstructor for the MIMIC-CXR dataset (Clark et al., 2019). We have split the training data in the ratio 4:1, and we used the greater subset as the training data and the smaller one as the validation data for the reconstructor, which is analogous to the approach applied with the JLiverCT dataset. We used binary cross-entropy loss to train the model, and applied Class Balanced Loss (CBL) (Cui et al., 2019) with $\beta = 0.999$. The number of parameters of the reconstructor was 110M. We fine-tuned the model with five epochs and conducted 5-fold cross-validation to determine the hyperparameters. The F-score on the validation dataset was 96.6. We used an Intel Core i9-9900K CPU and NVIDIA GTX 2080 GPU for training, and the training time was approximately 10 h.

C.3 Execution Time for Inference.

The execution time is crucial in the radiology report generation system for practical use. We calculated the execution time of our system trained with the JLiverCT dataset. In the end-to-end T5-Japanese-base model, the inference process incurred 0.4 seconds per report. However plain CoPlan model incurred approximately 10 seconds to conduct inference for one report. The slow inference process impedes the applicablity of the radiology report generation system.

To generate the reports faster with the CoPlan model, we employed several techniques in the inference phase. First, once the report evaluator quantified the report quality score of a plan, the estimated report quality score was cached to avoid recalculating the score. Radiology reports tend to be not very diverse in structure and sentence constructions. The same sentence structure, identified as a plan in our system, repeatedly appeared during the inference phase; therefore, caching the scores can drastically reduce the inference time of CoPlan. Second, the estimated quality scores $\text{RE}(\text{P}_{tg}(\hat{F}_{plan}))$ are updated only when the separate token of the plan ("<SEP>") is predicted. With these techniques, CoPlan performed inference for one report in 1.5 seconds.

C.4 Evaluation Settings.

Following (Dror et al., 2018), we use an approximate randomization test [16] to evaluate the statistical

[16] https://github.com/smartschat/art

	Correctness	Fluency	CO
JLiverCT Dataset (Time Series Critical Report)			
T5-Japanese-base	0.768	0.688	0.612
CoPlan	0.742	0.657	0.619
MIMIC-CXR Dataset (Time Series Irrelevant Report)			
4-class IC + T5-base	0.716	0.641	-
4-class IC + CoPlan	0.722	0.655	-

Table 12: Krippendorff's α for human evaluation on both MIMIC-CXR and JLiverCT datasets.

significance (sample size is 1,000). We calculated Krippendorff's alpha with the python Krippendorff library[17]. Table 12 shows Krippendorff's alpha scores for each metric on both the MIMIC-CXR and the JLiverCT datasets.

Evaluation Metrics on the JLiverCT Dataset. For the automatic evaluation of the JLiverCT dataset, we used BLEU (Papineni et al., 2002), F-scores of ROUGE-L (Lin, 2004), and CRS as metrics. We used the Natural Language Toolkit (NLTK) [18] to calculate the BLEU scores, and the ROUGE Python library [19] to calculate the ROUGE-L scores.

Evaluation Metrics on the MIMIC-CXR Dataset. For comparison with the previous image captioning approaches (Miura et al., 2021), we used BLEU-4 calculated by the NLTK library and BERTScore metrics (Zhang et al., 2019) [20] library. DistilBERT is used to calculate the BERTScore [21] aligning our experimental conditions with previous end-to-end research Miura et al. (2021). However, word-overlap-based metrics, such as BLEU, fail to assume the factual correctness of the generated reports. We compared the labels assigned in the CheXpert Labeler between the generated reports and gold reports to calculate the CheXpert accuracy, precision, micro F-score, and macro F-score. Note that we conducted a report-level evaluation in the same manner as for Miura et al. (2021), different from an image-level evaluation in Chen et al. (2020).

Details of the Annotators for the Human Evaluation. We outsourced a human evaluation task to the data annotation company with an adequate budget compared to the minimum wage in Japan. All six annotators for the MIMIC-CXR dataset and the two annotators for the JLiverCT dataset were Japanese but were also fluent in English and had substantial experience annotating medical corpora. Before requesting the evaluation task, we demonstrated an instruction for the human evaluation (Table 13) and agreed on the evaluation's purpose.

Instructions: We are currently working on a research project to automatically generate a report describing abnormalities and normalities in medical images, referred to as a "reading report." The goal is to determine whether the automatically generated report is of good or bad quality. We will demonstrate the automatic or human-generated reading reports.
We request to annotate the following two labels:

- **Content Evaluation** Annotate which findings are described and in which position?
- **Fluency Evaluation** Rate the fluency and readability of the report on a scale of 1 to 5.

Table 13: Instructions for the annotators on the human evaluation.

To evaluate the report in the MIMIC-CXR dataset, all annotators complete the "Data or Specimens Only Research" course of the CITI program [22] and received a certificate. This course deals with ethics of human subjects research and privacy-related matter to handle clinical datasets.

[17] https://github.com/pln-fing-udelar/fast-krippendorff

[18] https://www.nltk.org/

[19] https://github.com/pltrdy/rouge

[20] https://github.com/Tiiiger/bert_score

[21] distilbert-base-uncased_L5_no-idf_version=0.3.11(hug_trans=4.12.3)-rescaled

[22] https://www.citiprogram.org/