

Dynamic Traceback Learning for Medical Report Generation

Shuchang Ye, Mingyuan Meng, Mingjian Li, Dagan Feng, Usman Naseem, and Jinman Kim, *Member, IEEE*

Abstract—Automated medical report generation has the potential to significantly reduce the workload associated with time-consuming process of medical reporting. Recent generative representation learning methods have shown promise in integrating vision and language modalities for medical report generation. However, when trained end-to-end and applied directly to medical image-to-text generation, they face two significant challenges: i) difficulty in accurately capturing subtle yet crucial pathological details, and ii) reliance on both visual and textual inputs during inference, leading to performance degradation in zero-shot inference when only images are available. To address these challenges, this study proposes a novel multi-modal dynamic traceback learning framework (*DTrace*)¹. Specifically, we introduce a traceback mechanism to supervise the semantic validity of generated content and a dynamic learning strategy to adapt to various proportions of image and text input, enabling text generation without strong reliance on the input from both modalities during inference. The learning of cross-modal knowledge is enhanced by supervising the model to recover masked semantic information from a complementary counterpart. Extensive experiments conducted on two benchmark datasets, IU-Xray and MIMIC-CXR, demonstrate that the proposed *DTrace* framework outperforms state-of-the-art methods for medical report generation.

Index Terms—Report Generation, Cross-modal Generation, Medical Image Analysis, Multimodal Learning

I. INTRODUCTION

MEDICAL report generation is a crucial component of image diagnostic process, providing detailed textual descriptions of medical images to guide clinical decision-making and treatment planning [1]. However, the task of interpreting medical images and composing reports is both time-consuming and resource-intensive. Therefore, automating the report generation process, to produce a functional draft, has garnered significant attention as a potential solution to reduce radiologist workload [2]. Despite advances in deep learning, medical report generation remains challenging due to the difficulty in accurately associating subtle yet critical diagnostic features in medical images with their corresponding textual reports. This is because medical images often contain subtle regions, such as tumor areas, which are essential for diagnosis, while medical reports rely on a limited set of keywords to represent this diagnostic information. Hence, it is crucial to develop a method that can capture and associate this nuanced information to preserve the intrinsic medical meanings within medical images and reports.

Existing medical report generation methods usually rely on an encoder–decoder framework (Fig. 1a) that performs uni-modal learning to build a uni-directional mapping from images

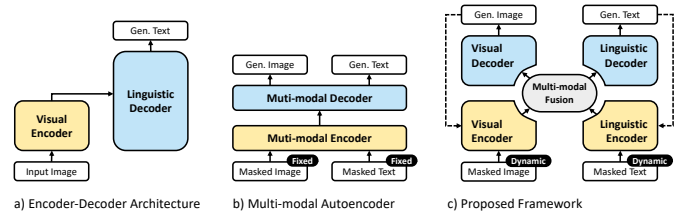


Fig. 1. Illustration of different generative frameworks. (a) Common uni-modal encoder-decoder framework, (b) Multi-modal masked encoder-decoder framework for generative representation learning (GRL), and (c) Our proposed framework with dynamic traceback learning (*DTrace*).

to reports [3]–[6]. While this imposes a forward semantic connection, it neglects the reciprocal relation that textual semantics map back to visual evidence. Without such bi-directional supervision, models often learn to exploit statistical regularities in the text (e.g., frequent phrases or template-like sentences) as shortcuts, rather than grounding their outputs in image evidence [7], as depicted in Fig. 2. To overcome this limitation, we argue that text-to-image generation is equally critical. By requiring text embeddings to reconstruct masked image regions, the model is compelled to encode language in a way that preserves pathology-relevant semantics. This bi-directional training enforces stronger cross-modal alignment, ensuring that generated reports are linguistically coherent and semantically grounded in the underlying visual content. Modeling such mutual associations between modalities has been shown to facilitate a more profound understanding of cross-modal knowledge [8]. Recently, Generative Representation Learning (GRL) methods have exploited multi-modal learning and generation [9]–[11], where image and report reconstruction are jointly performed in a multi-modal masked encoder-decoder framework (Fig. 1b). These methods aim to learn the latent space representations by masking inputs and subsequently reconstructing the original inputs based on the unmasked information. Although these methods have brought significant advancements in multi-modal learning, they exhibit considerable gaps when adopted for medical report generation due to two primary limitations:

First, GRL methods prioritize capturing superficial structural cues rather than deeper pathological semantics [12]. In images, this is reflected in morphological properties such as organ contours, shapes and spatial layouts, whereas in reports it manifests as structural regularities, including the templated phrasing and typical sentence organization of radiology narratives. These modality-specific surface features are often overemphasized, limiting the ability of GRL methods to

¹To facilitate reproducibility, we have made our code publicly available at: <https://github.com/ShuchangYe-bib/DTrace>

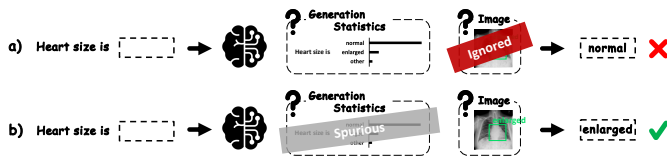


Fig. 2. Illustration of the limitation of existing report generation framework. a) The model make prediction based on spurious generation statistics, overlooking the radiology images; b) Ideally, the model should understand the pathological information in the image and generate a report accordingly.

represent fine-grained pathological semantics, such as lesion extent or disease progression. In image reconstruction, these methods often reduce the task to simple pixel matching by minimizing pixel-level discrepancies, neglecting the semantic and pathological nuances of images. For report reconstruction, GRL methods tend to predict frequently observed words to achieve a high overlap rate between the original and reconstructed reports. Due to the word imbalance in medical reports, where keywords signifying pathology appear infrequently, this learning approach potentially leads to the generation of clinically-flawed templated reports. Further, there exists inherent variability in the report descriptions for the same medical images, such as the order of reporting (e.g., starting from the disease sites or from image acquisition protocol) and the lexical choices employed to convey the severity of symptoms. However, the loss functions of GRL methods typically operate at the word level, which cannot measure the quality of generated reports from the sentence level with semantic contextual information.

Second, GRL methods focus on predicting masked images and text from their unmasked counterparts, which usually require a large amount of unmasked information to achieve text reconstruction [9], [11], with the mask ratio for the text being limited to a low range, $\approx 15\%$ (most textual information retained as the input). Whereas, the inference is based solely on images in medical report generation. Such a shift can lead to a performance drop (see Section V-C), as GRL methods are not designed to handle inference from images alone without accompanying text.

To address the above limitations, in this study, we propose a novel report generation framework to overcome the limitations of GRL methods and introduce dynamic traceback learning for medical report generation (DTrace). During inference, images are fully visible (0% mask ratio), and text is completely masked (100% mask ratio). During training, image and text mask ratios fluctuate between 0% and 100% per batch. We demonstrate that this variable masking approach is the most effective way of learning (see Section V-C). Our contributions are as follows:

- We introduce DTrace, a multi-modal framework for medical report generation that jointly learns bi-directional image-to-report and report-to-image generation, enhancing cross-modal knowledge by recovering masked semantic information.
- We introduce a traceback mechanism in DTrace that ensures semantic validity by reintegrating generated images and reports into their encoders for self-assessment.

- We introduce a dynamic learning strategy in DTrace that adapts to any image-text ratio, enabling effective training with both modalities and supporting image-only inference by adjusting loss weights dynamically.
- Extensive experiments on two benchmark datasets (IU-Xray and MIMIC-CXR) show that the proposed DTrace outperforms state-of-the-art medical report generation methods.

II. RELATED WORK

A. Medical Report Generation

Traditional medical report generation methods rely on rule- or template-based methods [13]. Rule-based methods often fall short in handling different scenarios and capturing language subtleties, while template-based methods are dependent on template quality and adaptability. With the paradigm shift in Computer Vision (CV) and Natural Language Processing (NLP), deep learning-based medical report generation methods have achieved promising performance and attained wide attention [1].

Deep learning-based report generation can be traced back to the encoder-decoder architecture, where images are transmuted into representative vectors through a visual encoder and then decoded into text [14]. Subsequent studies enhanced both components [15], evolving from CNN-RNN to Vision Transformer and Transformer [5], and later turned to improving interaction between the encoder and decoder. Visual language pretraining (VLP) has further advanced image-to-text generation by integrating visual and textual data; for example, MCGN [16] employs CLIP [17] with contrastive learning to refine latent spaces and boost semantic similarity. However, VLP approaches mainly strengthen encoders, while report generation also requires effective decoder training. In medical report generation, R2GenCMN [18] unified visual and linguistic representations with a shared vector pool, and XProNet [19] improved it using pathological labels, but both still relied on uni-directional image-to-report mapping. Recent methods incorporate additional annotations: COMG [20] uses segmentation masks to guide attention to critical regions, and RGRG [21] links bounding boxes with sentences to focus on specific anatomy. Others emphasize semantic alignment, such as ICT [22], which proposes a dual-level collaboration and inference network to enhance sentence coherence and report-level fidelity.

B. Generative Representation Learning

GRL methods learn the latent space representations by training the model to reconstruct the masked inputs based on the unmasked information. Masked Image Modeling (MAE) [10] and Masked Language Modeling (BERT) [9] are prevalent pre-training techniques in CV and NLP. BERT learned the word latent representations by training the model to predict the masked content based on the surrounding words. Then, MAE employed a similar strategy in images, where the images were split into patches and the model was trained to reconstruct the randomly masked patches. Recently, M3AE [11] exploited multi-modal learning and generation, where image and text

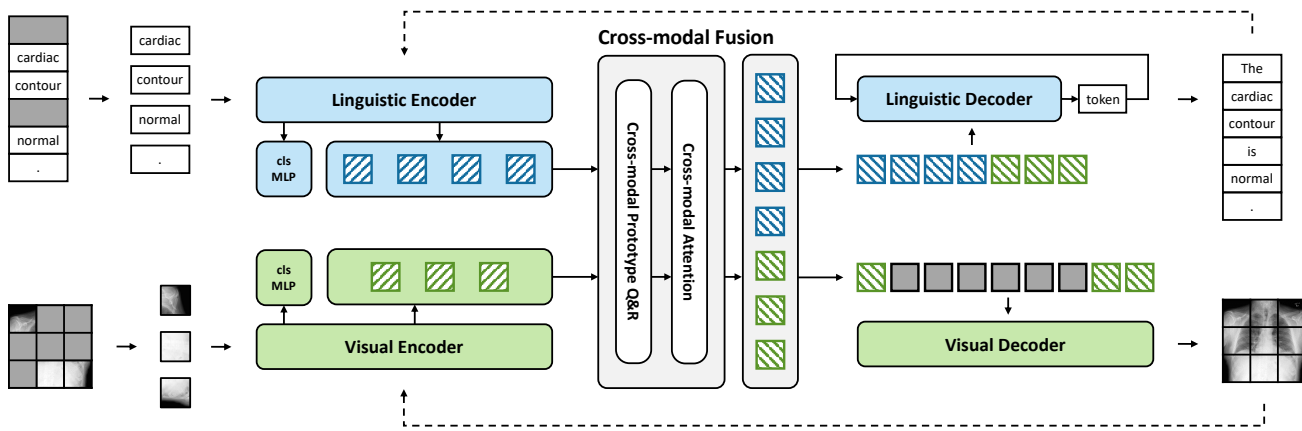


Fig. 3. DTrace with dynamic traceback learning. Solid and dashed lines indicate forward and traceback stages. Training: (1) Forward stage: masked images and masked reports are encoded by the visual and linguistic encoders, fused by the cross-modal module, and decoded to reconstruct images and reports. (2) Traceback stage: the reconstructed images and reports are re-encoded to verify semantic validity, with losses computed against unmasked ground-truth. Inference: the visual encoder processes the full image, and at each decoding step, the linguistic encoder encodes the autoregressively generated prefix tokens; fused features are passed to the linguistic decoder to predict the next report token until an end token is generated.

reconstruction were jointly performed with a multi-modal masked encoder-decoder framework to enhance the comprehension of cross-modal associations. These GRL methods were widely adopted as a pre-training step to enhance the performance of downstream tasks, such as disease classification [23] and medical visual question answering [24]. Subsequently, the concept of masking-based generative representation learning has been extended to various domains, including video understanding, as demonstrated by methods like MGMAE [25], and the enhancement of foundation model training, as exemplified by approaches such as EfficientSAM [26].

The capability of reconstructing masked text provides the potential for report generation. However, GRL methods were seldom applied to medical report generation due to the two drawbacks that we identified above. Recently, MedViLL [27] applied GRL framework to medical report generation (as one of the downstream tasks) by progressively replacing mask tokens with predicted language tokens. Unfortunately, its performance was limited (0.066 in BLEU4 in the MIMIC-CXR dataset) as it was not trained to handle situations where text information is not available.

In contrast, our proposed framework differs in several key aspects: i) GRL methods are typically designed for pre-training on large datasets followed by fine-tuning for specific tasks, whereas our framework is trained end-to-end and can be directly applied to medical report generation; ii) GRL usually employs a fixed mask ratio, but we utilize a dynamic mask ratio; and iii) while conventional methods prioritize morphological and structural similarity, our approach emphasizes semantic similarity.

III. METHOD

A. Network Architecture

DTrace consists of five key components: a visual encoder, a visual decoder, a linguistic encoder, a linguistic decoder, and a cross-modal fusion module (Fig. 3). The visual encoder processes partially masked images to extract incomplete pathological information. Simultaneously, the linguistic

encoder also processes fragmented textual information to extract incomplete pathological information. Then, the extracted information is fed into the cross-modal fusion module, where a cross-attention mechanism is employed to foster the interaction between the visual and linguistic domains, thereby ameliorating the semantic deficits in both modalities. After this, the enriched information is conveyed to the visual and linguistic decoders to restore the masked images and reports to their original unmasked states. Below, we discuss the key components of DTrace in detail:

Visual Encoder: consistent with the common practice of MAE; medical images were split into patches and then randomly masked. The visual encoder was a standard ViT [28], which mapped the unmasked patches into latent representations and then performed multi-label classification to predict the disease labels extracted via CheXbert [29].

Linguistic Encoder: medical reports were mapped into embedded text tokens following the standard pre-processing steps [9]. Then, these tokens were randomly masked and fed into the linguistic encoder. The linguistic encoder was a classic text transformer block [3].

Cross-Modal Fusion Module: the features extracted from both the encoders were projected to a pre-defined dimension, which then was then concatenated and fed to a cross-modal attention module [11] for information interchange. The resultant features were subsequently separated and mapped into latent representations via Multi-layer Perceptrons (MLPs) [30].

Visual Decoder: the masked tokens were reinstated to their original positions, aligning with the unmasked encoding tokens. Following this, a lite version of ViT [10] was used to restore the masked patches.

Linguistic Decoder: a relational memory module is incorporated to capture recurrent report patterns across similar images. At each decoding step t , the memory matrix from the previous step (M_{t-1}) serves as the query, while its concatenation with the previous output embedding y_{t-1} serves as the key and value for the multi-head attention. The updated memory M_t is further refined through residual connections

Algorithm 1 Dynamic Traceback Learning

```

1: Generate random numbers:  $\alpha$  as image mask ratio and
    $\beta = 1 - \alpha$  as text mask ratio
2: while Not reach max epoch and early stop do
3:   if Forward then
4:     imasked, tmaskd  $\leftarrow$  masking(image, text)
5:     ifeats, tfeats  $\leftarrow$  encoders(imasked, tmaskd)
6:     compute  $\ell_{FVD}$  and  $\ell_{FLD}$ 
7:     gradient descent encoders by  $(1 - \alpha) \cdot \ell_{FVD}$  and  $(1 - \beta) \cdot \ell_{FLD}$ 
8:     feats  $\leftarrow$  cross modal fusion(ifeats, tfeats)
9:     igen, tgen  $\leftarrow$  decoders(feats)
10:    compute  $\ell_{IR}$  and  $\ell_{RG}$ 
11:    gradient descent all components by  $\alpha \cdot \ell_{IR}$  and  $\beta \cdot \ell_{RG}$ 
12:  end if
13:  if Traceback then
14:    disable the gradient descent of encoders
15:    igfeats, tgfeats  $\leftarrow$  encoders(igen, tgen)
16:    compute  $\ell_{TVD}$  and  $\ell_{TLD}$ 
17:    gradient descent all except encoders by  $\alpha \cdot e^{-\ell_{FVD}} \ell_{TVD}$ 
      and  $\beta \cdot e^{-\ell_{FLD}} \ell_{TLD}$ 
18:  end if
19: end while

```

and gated mechanisms to stabilize gradient propagation.

B. Traceback Mechanism

The traceback mechanism (Fig. 4) was developed to ensure the medical validity of generated outputs through two phases: forward and traceback. Initially, encoder capabilities are enhanced to identify pathological information during the forward phase. Subsequently, the decoder outputs are redirected to their corresponding encoders in the traceback stage to check their medical validity. Through repetitive iterations between the two stages, the DTrace model enhances accuracy by refining its understanding of medical content, ensuring reliable identification and validation of pathological information in image reconstruction and report generation.

C. Dynamic Learning Strategy

Existing multi-modal GRL methods incurred performance degradation when generating reports from images alone. To address this limitation, the dynamic learning strategy was proposed to enhance the generalizability of the model to perform multi-modal generation given any percentage of text and

image inputs. This was achieved through training with various complementary image and text mask ratios (different mask ratios for each batch). This complementary relation aimed to guarantee sufficient shared information and to ensure that each modality can consistently derive some information from the other. It also incorporated a self-adjustment mechanism for loss weights, dynamically adapting to changes in the mask ratios. The pseudo-code of the dynamic traceback learning is shown in Algorithm 1. In the subsequent sections, the mathematical rationale behind the adjustment of loss weights during the forward and traceback stages is presented.

Assumed the mask ratio of images to be a random number $0 \leq \alpha \leq 1$. The corresponding mask ratio of reports would then be $\beta = 1 - \alpha$. Notably, when images are heavily masked, retaining informative textual tokens becomes crucial, as they provide semantic anchors that enrich morphological recovery with diagnostic meaning; conversely, when text is sparsely available, visual morphology compensates, ensuring balanced cross-modal fusion. Although complementary mask ratios do not strictly guarantee non-overlapping information, the stochastic token- and patch-level masking in DTrace ensures that, across training iterations, both modalities encounter diverse missing-information scenarios. This randomness, when combined with semantic supervision from the traceback stage, encourages robust cross-modal compensation even in the presence of partial redundancy [9]–[11], thereby enabling effective information transfer between modalities.

Forward Stage The forward stage of the DTrace model performs three main tasks: 1) disease identification by the encoders, 2) image reconstruction by the visual decoder, and 3) report generation by the linguistic decoder, as shown in the left of Fig. 4.

For disease identification, we integrated a multi-label classification head for N classes into the encoders to predict disease labels as output. According to the dynamic learning strategy, the image and reports are masked in the forward stage based on the dynamic complementary mask ratio. Subsequently, the remaining unmasked components are directed to their respective encoders. To optimize the generation of their corresponding label outputs, both the encoders aim to minimize the diagnostic loss. The diagnostic loss is the binary cross-entropy (*BCE*) loss (see Equation 1) between the instances of disease label extracted by CheXbert, y and the encoders' corresponding predicted label, \hat{y} . As a diagnosis can only be meaningful when sufficient unmasked information is available, the visual and linguistic encoders' respective diagnostic losses complement their corresponding image and text mask ratios. The resulting forward visual diagnostic (*FVD*) loss and forward linguistic diagnostic (*FLD*) loss are shown in Equation 2 and Equation 3.

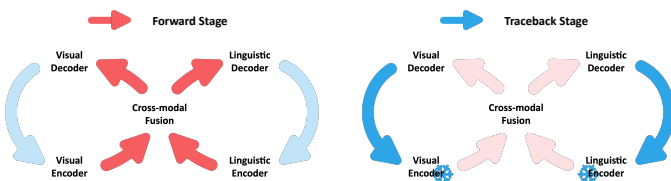


Fig. 4. Dynamic traceback learning with forward stage (left) and traceback stage (right). The freeze means that there is no gradient descent in back-propagation.

$$\ell_{BCE}(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^N y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i) \quad (1)$$

$$\ell_{FVD}(y, \hat{y}) = (1 - \alpha) \cdot \ell_{BCE}(y, \hat{y}) \quad (2)$$

$$\ell_{FLD}(y, \hat{y}) = (1 - \beta) \cdot \ell_{BCE}(y, \hat{y}) \quad (3)$$

For image reconstruction, we follow the MAE's practice [10] in that the visual decoder takes the visual features that gained knowledge from the cross-model fusion module and learned vector that indicated the presence of the masked patch as input. The image reconstruction process is regulated through minimizing the image reconstruction (*IR*) loss, which is the pixel-wise mean-square-error (*MSE*) weighted proportional to the image mask ratio to facilitate cross-modal communication, as shown in Equation 4. In Equation 4, g represents an instance of the origin image and \hat{g} represents the respective reconstructed image, and W and H represent the width and height of the image, respectively.

$$\ell_{IR}(g, \hat{g}) = \alpha \cdot \frac{1}{W} \frac{1}{H} \sum_{i=1}^W \sum_{j=1}^H (g_{ij} - \hat{g}_{ij})^2 \quad (4)$$

The generated reports are refined by minimizing the report generation (*RG*) loss. The *RG* loss, as shown in Equation 5, is the word-level cross-entropy (CE) loss weighted corresponding to the text mask ratio to attend to instances where reports are generated solely from images, where r represents an instance of ground truth report, \hat{r} represents the corresponding generated report, V represents the vocabulary size and L represents the report length.

$$\ell_{RG}(r, \hat{r}) = \beta \cdot \frac{1}{V} \frac{1}{L} \sum_{i=1}^V \sum_{j=1}^L r_{ij} \cdot \log\left(\frac{e^{\hat{r}_{ij}}}{\sum_{k=1}^V e^{\hat{r}_{ik}}}\right) \quad (5)$$

Traceback Stage As depicted in the right of Fig. 4, the images reconstructed and reports generated by the decoders were traced back to be the input of encoders with frozen parameters. The encoders inference on both the reconstructed items (\hat{y}) and unmasked items (\tilde{y}). To avoid error accumulation when reconstructed outputs are fed back into the encoders, traceback losses are inversely scaled by forward errors, such that unreliable reconstructions are down-weighted while accurate outputs progressively gain influence, ensuring stable and semantically valid feedback, as shown in Equation 6 and Equation 7. As the forward errors decrease and reconstructions become more accurate, the traceback losses naturally gain more weight, thereby progressively reinforcing semantic consistency without compounding errors.

$$\ell_{TVD}(\tilde{y}, \hat{y}) = \alpha \cdot e^{-\ell_{FVD}(\tilde{y}, \hat{y})} \cdot \ell_{BCE}(\tilde{y}, \hat{y}) \quad (6)$$

$$\ell_{TLD}(\tilde{y}, \hat{y}) = \alpha \cdot e^{-\ell_{FLD}(\tilde{y}, \hat{y})} \cdot \ell_{BCE}(\tilde{y}, \hat{y}) \quad (7)$$

IV. EXPERIMENTAL SETUP

A. Dataset

We conducted experiments on two well-benchmarked public datasets: Indiana University Chest X-ray (IU-Xray) [31] and MIMIC Chest X-ray (MIMIC-CXR) [32]. We split the data into training, validation, and testing subsets. For IU-Xray, we adopt the widely accepted 7:1:2 data split as suggested in prior studies [4], [18], [19]. For MIMIC-CXR, we adhere to the official data split.

B. Evaluation Metrics

To measure the quality of generated medical report, we follow the standard practice [4], [5], [18], [19] to adopt natural language generation (NLG) metrics: BLEU [33], METEOR [34], ROUGE-L [35], and CIDEr [36] as the evaluation metrics. We further assess the clinical efficacy (CE) [4], [18], [37], [38] of the generated reports by annotating them with CheXbert and comparing the predicted and ground truth labels.

C. Implementation Details

Our implementation of DTrace was realized using the PyTorch package [39]. Optimization of the gradient descent process was carried out utilizing the AdamW optimizer [40], with a set learning rate of 10^{-4} . In the context of report generation, the beam search algorithm was employed, with a specified beam width of 3. The predefined maximum lengths for the reports were set at 60 and 100 for the IU-Xray and MIMIC-CXR datasets, respectively. Training of the model was performed on an NVIDIA RTX A6000 graphics card, with a designated mini-batch size of 16. To enhance the efficiency of training, the model initially underwent a pre-training phase, wherein it was conditioned to reconstruct images and reports based on the information inherent to their respective modalities, coupled with contrastive learning techniques.

All reported results and discussions in this manuscript are based on an image-only inference setting, in which all text inputs are fully masked during testing to rigorously evaluate the model's ability to generate reports solely from image inputs. In terms of real-world deployment, the deployed DTrace model has a size of 1,324.74 MB, an average inference latency of 755.77 ms per batch, and an inference throughput of 1.32 FPS.

D. Baselines

We compared the performance of DTrace against the state-of-the-art medical report generation methods (see Section V-A): R2Gen [4], R2GenCMN [18], CMCL [41], Align-Transformer [6], TriNet [42], MCTransformer [5], M2KT [43], ORGAN [37], KiUT [38], RAMT [44], pretrain+FMVP [45], Med-LLM [46], MultiP [47] and CK-Net [48]. For fair evaluation, the released code from the baseline methods was used in the same settings as described in the papers. Additionally, we included methods such as COMG [20], oracle+FMVP [45], and RGRG [21], which leverage auxiliary annotations beyond textual reports, to provide a holistic benchmarking comparison. To further analyze DTrace's capabilities, we performed an ablation study to evaluate the contributions of individual components. To quantify the traceback paradigm's unique advantages against reconstruction-based approaches, we compare our Dtrace against M3AE [11], which reconstructs masked image patches and report tokens using standard reconstruction losses only, without applying traceback or weighted diagnostic supervision. We also investigated the impact of varying the mask ratio on model performance during training. Finally, we present qualitative comparisons between the reports generated by DTrace and those produced by competing methods, offering insights into the semantic and contextual differences.

TABLE I

PERFORMANCE COMPARISON BETWEEN DTRACE AND EXISTING REPORT GENERATION METHODS ON THE IU-XRAY AND MIMIC-CXR DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD AND THE SECOND BEST ARE UNDERLINED. BL, MTR, RG-L AND CDR ARE THE ABBREVIATIONS OF NLG EVALUATION METRICS BLEU, METEOR, ROUGE-L AND CIDER. P, R, F ARE THE ABBREVIATIONS OF CE METRICS: PRECISION, RECALL, F1-SCORE. GRAY INDICATES THE UTILIZATION OF ADDITIONAL ANNOTATIONS. * DENOTES THAT THE RESULTS ARE CITED FROM THEIR ORIGINAL PAPERS.

Dataset	Model	BL-1	BL-2	BL-3	BL-4	MTR	RG-L	CDr	P	R	F
IU-Xray	R2Gen	0.470	0.304	0.211	0.157	0.197	0.364	0.342	-	-	-
	R2GenCMN	0.486	0.307	0.216	0.156	0.212	0.374	0.331	-	-	-
	CMCL*	0.473	0.305	0.217	0.162	0.186	0.378	-	-	-	-
	AlignTransformer*	0.484	0.313	0.225	0.173	0.204	0.379	-	-	-	-
	TriNet*	0.478	0.344	0.248	0.180	-	0.398	0.439	-	-	-
	MCTransformer*	0.496	0.319	0.241	0.175	-	0.377	<u>0.449</u>	-	-	-
	M2KT*	0.497	0.319	0.230	0.174	-	0.399	<u>0.407</u>	-	-	-
	ORGAN*	0.494	0.335	0.247	<u>0.190</u>	0.203	<u>0.395</u>	-	-	-	-
	KiUT	0.525	0.360	<u>0.251</u>	<u>0.185</u>	0.242	0.409	-	-	-	-
	RAMT	0.482	0.310	0.221	0.165	0.195	0.377	-	-	-	-
	Pretrain+FMVP	0.485	0.315	0.225	0.169	0.201	0.398	-	-	-	-
	MedLLM	-	-	-	0.168	0.209	0.381	0.427	-	-	-
	MultiP	0.515	0.332	0.238	0.178	0.216	0.386	-	-	-	-
	CK-Net	0.503	0.325	0.236	0.179	0.202	0.373	-	-	-	-
	DTrace (ours)	<u>0.516</u>	<u>0.353</u>	0.278	0.204	<u>0.233</u>	<u>0.386</u>	0.469	-	-	-
Oracle+FMVP*	0.499	0.339	0.256	0.206	0.211	0.422	-	-	-	-	
COMG*	0.482	0.316	0.233	0.184	0.191	0.382	-	-	-	-	
MIMIC-CXR	R2Gen	0.344	0.208	0.140	0.100	0.135	0.271	0.146	0.333	0.273	0.276
	R2GenCMN	0.327	0.211	0.148	0.109	0.137	<u>0.298</u>	0.135	0.334	0.275	0.278
	CMCL*	0.344	0.217	0.140	0.097	0.133	0.281	-	-	-	-
	AlignTransformer*	0.378	0.235	0.156	0.112	0.158	0.283	-	-	-	-
	TriNet*	0.362	0.251	0.188	0.143	-	0.326	0.273	-	-	-
	MCTransformer*	0.351	0.223	0.157	0.118	-	0.287	<u>0.281</u>	-	-	-
	M2KT*	0.386	0.237	0.157	0.111	-	0.274	<u>0.111</u>	-	-	-
	ORGAN*	0.405	<u>0.254</u>	<u>0.170</u>	<u>0.121</u>	<u>0.161</u>	0.291	-	0.416	<u>0.418</u>	<u>0.385</u>
	KiUT	0.393	0.243	0.159	0.113	0.160	0.285	-	0.371	0.318	0.321
	RAMT	0.362	0.229	0.157	0.113	0.153	0.284	-	0.380	0.342	0.335
	Pretrain+FMVP	0.389	0.236	0.156	0.108	0.150	0.284	-	-	-	-
	MedLLM	-	-	-	0.128	0.161	0.289	0.265	0.294	0.277	0.281
	CK-Net	0.374	0.235	0.157	0.116	0.150	0.284	-	0.386	0.379	0.384
	DTrace (ours)	<u>0.392</u>	0.260	0.171	0.129	0.162	0.309	0.311	<u>0.411</u>	0.436	0.391
	Oracle+FMVP*	0.391	0.249	0.172	0.125	0.160	0.304	-	-	-	-
COMG*	0.346	0.216	0.145	0.104	0.137	0.279	-	0.424	0.291	0.345	
RGRG	0.373	0.249	0.175	0.126	0.168	0.264	0.495	0.461	0.475	0.447	

V. RESULT AND DISCUSSION

A. Comparisons to Previous Methods

Table I presents a comprehensive comparison with state-of-the-art methods, highlighting the superior performance of the proposed DTrace across both NLG and clinical efficacy (CE) metrics on the IU-Xray and MIMIC-CXR datasets. DTrace consistently ranks among the top performers in NLG metrics, particularly excelling in BLEU-3, BLEU-4, and CIDEr, which underscores its strong capability in generating structurally accurate and linguistically coherent reports. In terms of CE metrics, DTrace achieves the best results in recall and F1-score and the second-best in precision, demonstrating its effectiveness in capturing disease regions and generating semantically meaningful reports. Even in comparison with models like COMG and RGRG, which leverage additional segmentation masks and region-specific information, DTrace remains highly competitive.

This superior performance suggests that DTrace is adept at capturing pathology-critical information, making it well-suited for scenarios where radiologists' descriptions of the same radiology image vary in writing style and terminology. For example, when comparing the ground truth sentence "the heart size is normal" with two variants, "the heart size is enlarged"

and "the heart size is within normal limits," the cross-entropy is smaller for the first variant. To address this potential bias towards common phrasing, which might sacrifice semantic accuracy, we introduce a traceback mechanism. This mechanism evaluates generated reports using an encoder trained for accurate diagnosis, thus reinforcing semantic correctness and diluting the impact of cross-entropy loss.

The proposed DTrace outperforms models such as R2GenCMN, AlignTransformer, and XProNet, which we attribute to the benefits of multi-modal learning, particularly in facilitating cross-modal communication. The traceback mechanism further contributes to this performance improvement. Even when a modality lacks information from another, it can reconstruct its form independently, while the traceback mechanism ensures that semantic information is still obtained from another modality. In the cross-modal fusion module, different modalities exchange and complement each other's information, thereby establishing a robust communication protocol. This approach achieves effects similar to R2GenCMN and XProNet, with the added enhancement of the traceback mechanism, leading to overall improved performance.

TABLE II
ABLATION STUDY ON KEY COMPONENTS OF OUR PROPOSED DTRACE ON MIMIC-CXR DATASET.

Model	BL-1	BL-2	BL-3	BL-4	MTR	RG-L	CDr	P	R	F
Uni-modal Report Auto-completion	0.351	0.215	0.141	0.102	0.133	0.264	0.108	0.277	0.244	0.236
Encoder-Decoder (Baseline)	0.348	0.212	0.143	0.106	0.136	0.277	0.143	0.325	0.271	0.268
+ Bi-directional Generation	0.346	0.220	0.144	0.107	0.142	0.285	0.241	0.345	0.280	0.280
+ Dynamic Learning	0.371	0.243	0.165	0.120	0.155	0.281	0.279	0.358	0.355	0.344
+ Traceback Mechanism	0.392	0.260	0.171	0.129	0.162	0.309	0.311	0.411	0.436	0.391
Multi-modal Masked Autoencoder	0.364	0.246	0.164	0.119	0.153	0.284	0.292	0.354	0.296	0.301

TABLE III
COMPARISON OF THE TRADITIONAL FIXED MASK RATIO TRAINING STRATEGY VERSUS OUR DYNAMIC LEARNING STRATEGY ON THE IU-XRAY DATASET, WITH OUR METHOD HIGHLIGHTED IN BOLD.

Strategy	Mask Ratio	BL-1	BL-2	BL-3	BL-4	MTR	RG-L	CDr
Fixed	0%	0.397	0.255	0.183	0.139	0.164	0.367	0.364
	15%	0.441	0.284	0.207	0.158	0.180	0.378	0.427
	30%	0.459	0.293	0.207	0.155	0.176	0.368	0.255
	45%	0.465	0.299	0.211	0.155	0.176	0.367	0.289
	60%	0.468	0.300	0.214	0.159	0.196	0.382	0.308
	75%	0.479	0.307	0.221	0.164	0.195	0.385	0.364
Dynamic	varying	0.516	0.353	0.278	0.204	0.233	0.386	0.469

B. Ablation Study

The ablation study results presented in Table II demonstrate the contribution of each individual component to the overall performance of DTrace. Below, we provide a concise discussion of the impact of each component of DTrace.

Bi-directional Multi-modal Generation: Incorporating bi-directional generation significantly improved the model’s ability to capture mutual associations between medical images and their corresponding reports. This enhancement led to a notable increase in CIDEr, from 0.143 to 0.241, and a slight improvement in the F1 score to 0.280. The bi-directional nature of the approach allows the model to better align the content of the generated reports with visual features, thereby improving both the relevance and accuracy of the textual output.

Dynamic Learning: The introduction of dynamic learning resulted in performance gain across all metrics, particularly in BLEU-4, which increased from 0.107 to 0.120, and the F1 score, which rose to 0.344. The dynamic learning strategy compels each modality not only to reconstruct its own content but also to rely on the complementary modality for missing semantics. As a result, the model simultaneously learns stronger unimodal features and more aligned visual–language features, enhancing cross-modal fusion’s effectiveness. The observed gains in NLG metrics indicate that the model is no longer confined to reproducing surface-level regularities or statistical patterns in the training corpus. Instead, it develops a stronger perception ability to capture salient pathological semantics. In parallel, the improvements in CE metrics demonstrate that the generated contents are more clinically aligned with radiological findings, reflecting that the strategy enhances linguistic fluency and ensures that the generated reports faithfully convey medically relevant information.

Traceback Mechanism: The traceback mechanism provided

the most substantial improvements. By supervising the semantic validity of the generated content, this component enhanced the model’s ability to produce clinically accurate and coherent reports. The BLEU-4 score increased to 0.129, while CIDEr reached 0.311, indicating improved alignment with the true medical meaning of the reports. Clinically, the model achieved a precision of 0.411, recall of 0.436, and an F1 score of 0.391. These improvements underscore the traceback mechanism’s critical role in ensuring that the generated text not only follows the expected structure but also conveys accurate and meaningful clinical information. Notably, while the reconstruction-only multimodal autoencoder baseline achieves moderate results, its performance is consistently inferior to our traceback design. This shows that reconstruction alone primarily captures surface-level morphological or lexical regularities, but fails to guarantee semantic correctness. In contrast, the traceback mechanism reintegrates generated outputs into the encoders, compelling the model to verify and refine pathological semantics.

Understanding of Cross-modal Knowledge: The DTrace framework demonstrates clear superiority over both the baseline Encoder-Decoder model and the Multi-modal Masked Autoencoder. While the baseline model struggles to generate clinically accurate reports, often relying on statistical regularities, DTrace’s dynamic learning strategy and traceback mechanism ensure that the generated content is both semantically valid and clinically meaningful. Compared to the Multi-modal Masked Autoencoder, which integrates vision and language modalities but faces challenges in maintaining semantic precision, DTrace excels by dynamically adapting to varying input proportions and supervising the clinical relevance of the generated content.

C. Impact of Mask Ratios

To illustrate the significance of dynamic learning in medical report generation, we compared traditional methods that

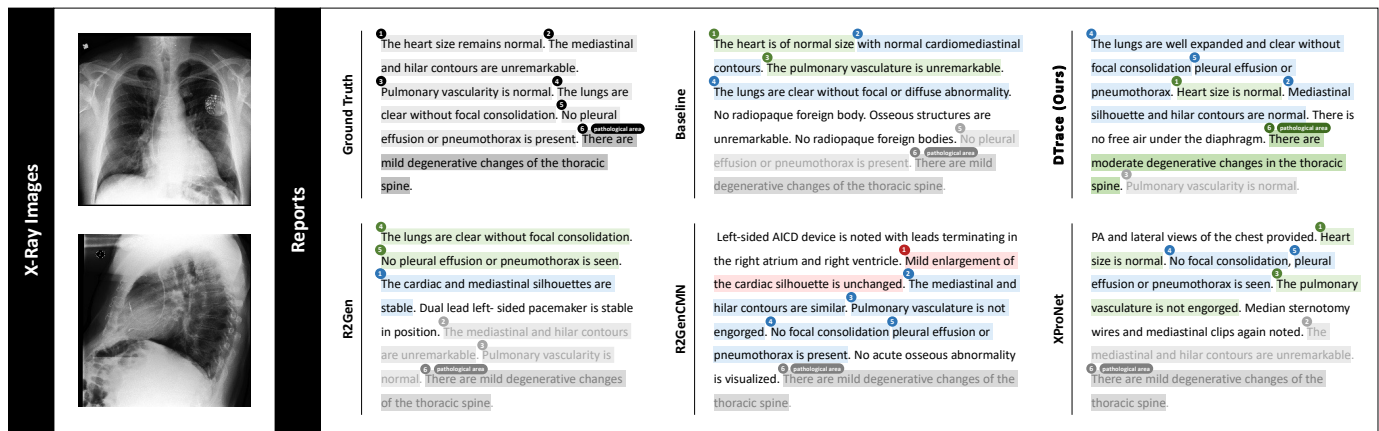


Fig. 5. An example (patient 10014765) of comparisons between different report generation frameworks and the proposed DTrace framework. The information in the ground truth report is labeled from 1 to 6 and highlighted separately. The generated reports are labeled according to the ground truth report and highlighted with different colors to represent the differences between the generated sequences and the ground truth report: (1) Green - consistent; (2) Blue - semantically similar but different in expression; (3) Pink - incorrect information; (4) Gray - missing sentences; (5) Unhighlighted - not included in the ground truth.

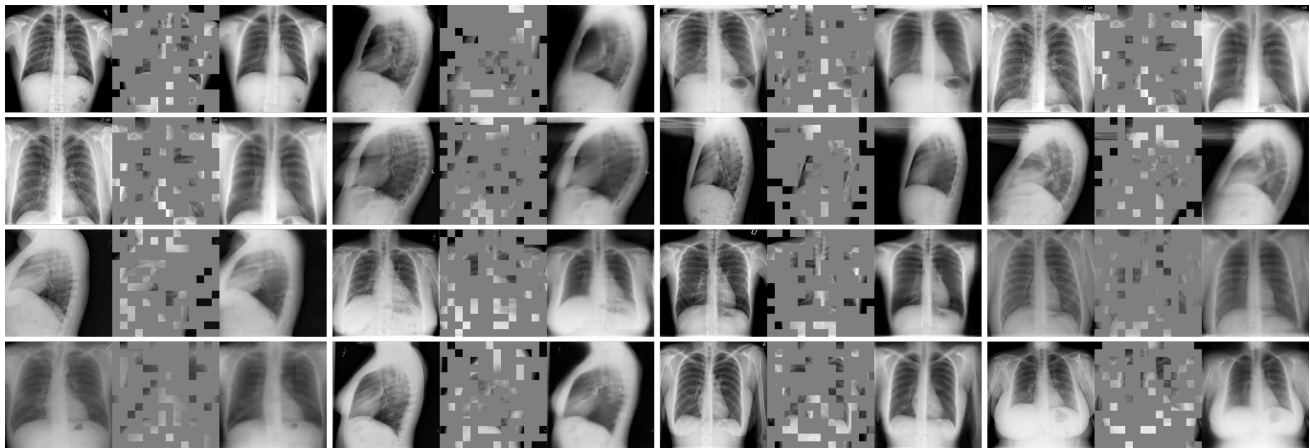


Fig. 6. Visualization of the reconstructed images with a mask ratio of 75% from the model. For each triplet, we show the original image (left), the masked image (middle), and our reconstructed image (right).

use a fixed mask ratio with our proposed method, which employs a varying mask ratio, as shown in Table III. Our comparison revealed that the model’s performance improves as the mask ratio increases. Importantly, during training, the model is exposed to both image and text data, whereas, during inference, it processes only the image data. The greater the discrepancy between the information available during training and inference, the worse the model’s performance tends to be. Our dynamic learning approach effectively mitigates this issue by randomly varying the proportions of information from both modalities, thus enhancing the model’s robustness during inference.

D. Qualitative Analysis and Visualization

We conducted a qualitative analysis of DTrace compared to the baseline encoder-decoder framework. As illustrated in Fig. 5, when dealing with rare diseases, the baseline and existing methods frequently omit critical diagnostic statements, such as “mild degenerative changes in the thoracic spine,” despite having a high degree of textual overlap with the

	Original	Masked	Without Traceback	With Traceback
Frontal				
Lateral				
			FID	
			166.4	98.6

Fig. 7. Comparison of visualization of reconstructed images with and without dynamic traceback learning.

ground truth. In contrast, the DTrace-generated reports include most of the essential diagnostic statements and demonstrate a high level of consistency with the ground-truth reports.

TABLE IV
THE CLASSIFICATION PERFORMANCE COMPARISON BETWEEN ORIGINAL IMAGES AND RECONSTRUCTED IMAGES.

	Accuracy	Precision	Sensitivity
Original	0.861	0.922	0.904
Reconstructed	0.868	0.909	0.931

Additionally, Fig. 6 presents images reconstructed from 75%-masked images and unmasked reports by DTrace. These reconstructed images exhibit a high degree of consistency with the original unmasked images, highlighting DTrace’s effectiveness in image reconstruction. Further visualizations can be found in the supplementary materials.

We further assessed the morphological and semantic similarities between the constructed images and the original images. To evaluate the efficacy of our dynamic traceback learning in reducing pixel-level differences, we conducted a control experiment comparing the quality of images generated with and without dynamic traceback learning, as depicted in Fig. 7. Our analysis revealed a distinct boundary between generated patches and original patches in images not reconstructed using the dynamic traceback learning strategy. Although dynamic traceback learning is not specifically designed to enhance morphological similarity, the images reconstructed using this method appear more cohesive and clearer.

To quantify the quality of the generated images, we compared the Frchet Inception Distance (FID) scores. After implementing dynamic traceback learning, the FID of the constructed images decreased from 166.4 to 98.6, indicating that dynamic traceback learning more accurately mimics the distribution of the original images and produces higher-quality, more realistic reconstructions. Additionally, we employed a classification evaluation method to assess the semantic correctness of the reconstructed images. Table IV demonstrates that DTrace successfully reconstructs images from 75%-masked images and unmasked text input while preserving semantic information.

E. Interpretability

To deeper understand the association between image reconstruction and report generation, we conducted an inter-

pretability analysis by visualizing how reconstructed images affect report generation. Figure 7 illustrates two representative cases, demonstrating how the model leverages reconstructed visual semantics to influence textual outputs. In the left case (CXR145_IM-0290), the ground truth report states, “In the left lower lobe, a patchy infiltrate is present.” The generated report correctly describes, “Lungs demonstrate left lobe airspace disease most consistent with pneumonia.” Although the terminology differs, these descriptions are clinically consistent. In radiologic interpretation, a patchy infiltrate refers to irregular, scattered areas of increased lung opacity, suggesting partial filling of the alveoli by fluid, pus, blood, or cells, without a uniform or lobar distribution. Conversely, lobe airspace disease denotes a more localized and confluent filling of alveoli within a single lung lobe, as seen in lobar pneumonia. When patchy infiltrates are predominantly confined to one lobe and demonstrate airspace-filling features, they are synonymous with lobe airspace disease. Thus, the model’s output captures the same underlying pathology despite the lexical variation. The reconstructed image and highlighted region of interest also exhibit strong alignment with the original abnormality, indicating that the traceback mechanism preserves semantic information critical for accurate report generation. In contrast, the right case (CXR64_IM-2218) illustrates a failure scenario. The ground truth report notes “Heart size is mildly enlarged,” yet the generated report incorrectly states, “The heart is normal in size.” Here, the reconstructed image erroneously depicts the heart as normal, leading to a semantic mismatch between visual reconstruction and textual output. This example highlights an important limitation: when the reconstruction of a key anatomical structure is inaccurate, errors can be propagated into the generated report. Collectively, these cases demonstrate that reconstructed images function as a transparent window into the model’s decision-making process, revealing both its ability to preserve nuanced semantic consistency and its vulnerability when foundational visual cues are misrepresented.

F. Learning Process

Our investigation further delved into the model’s learning process by visualizing the evaluation metrics of the validation set as a function of epochs, as depicted in Fig. 9. We noted

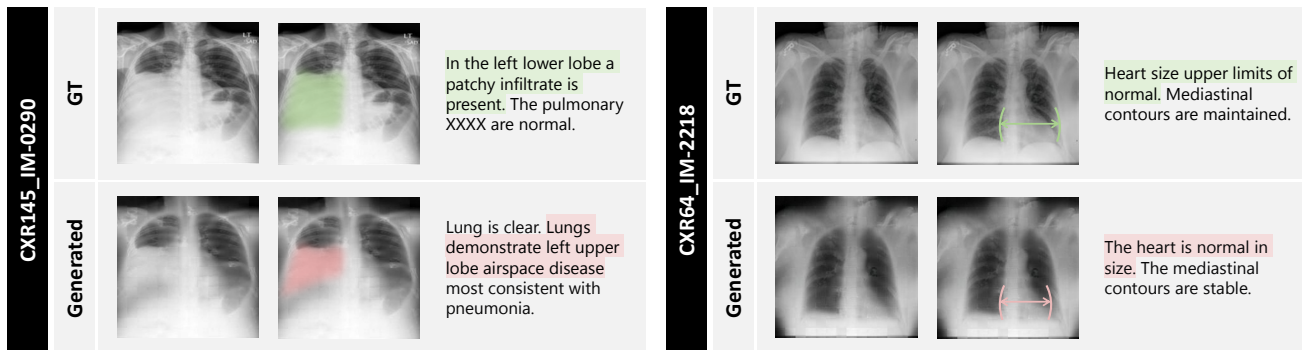


Fig. 8. Interpretability analysis of DTrace. For each case, the upper row presents the ground-truth image and reference report, while the lower row shows the corresponding outputs generated by DTrace. Within each row, we display the original image (left), the highlighted region of interest (middle), and the associated report (right).

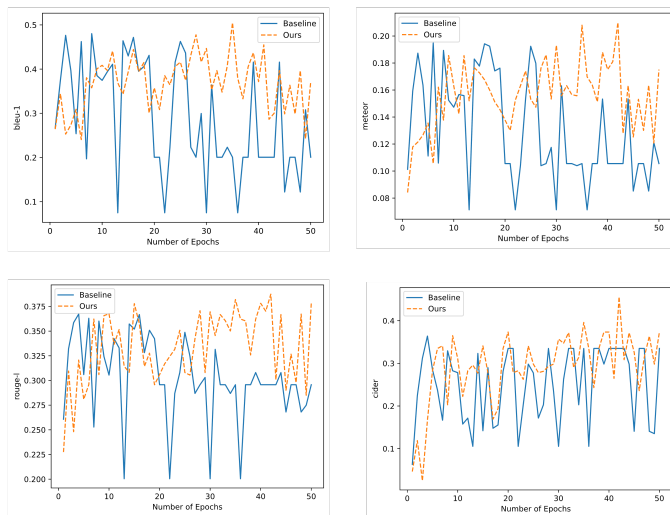


Fig. 9. Comparative Visualization of Learning Process. The x-axis represents the epoch count, while the y-axis quantifies evaluation metrics (BLEU, METEOR, ROUGE-L, and CIDEr). The baseline model, featuring a traditional encoder-decoder architecture, is delineated by the blue line, whereas our novel framework is depicted by the orange line.

that traditional models for report generation exhibited considerable fluctuations in performance metrics during training and were prone to entrapment in local minima. When trapped in a local minimum, the model tended to produce “average reports,” generating identical reports for any input image, which corresponded to a relatively low cross-entropy loss. However, such models are practically futile, as this is a consequence of data imbalance and inconsistency in report expression. In cases where descriptions pertain to a specific organ, the majority of instances are deemed normal, leading the model to a complacency that precludes learning how to extract features and transmit visual information to the text generator. To minimize cross-entropy loss, the model might adopt a uniform expression approach. In short, traditional methods, in our context, incline the model to learn a template that minimizes cross-entropy loss, rather than diagnosing the radiology images. By incorporating a traceback mechanism and a dynamic learning strategy, our model is compelled to engage in exchange and extract semantic information. Revealing parts of the report apprises the model of the general expression style of the input, thereby concentrating its focus on the alignment of key information. Consequently, this approach results in smaller fluctuations in performance metrics during training and enhances overall performance.

VI. LIMITATIONS

The primary limitation of this study is the elevated computational cost incurred by the additional image decoder and text encoder during the training phase. However, once the multi-modal generation model is fully trained, it becomes feasible to employ pruning techniques to remove components that are irrelevant to specific tasks, such as report generation in our scenario.

VII. CONCLUSION

In this study, we introduced a novel medical report generation framework, DTrace, which leverages multi-modal dy-

amic traceback learning. We incorporated a traceback mechanism to ensure semantic correctness during training and a dynamic learning strategy to reduce the dependency of existing generative cross-modal frameworks on textual input. Experimental evaluations demonstrate that the incorporation of these components significantly enhances the multi-modal generation capabilities of the framework, enabling DTrace to achieve state-of-the-art performance on the benchmark IU-Xray and MIMIC-CXR datasets.

REFERENCES

- [1] P. Messina, P. Pino, D. Parra, A. Soto, C. Besa, S. Uribe, M. Andía, C. Tejos, C. Prieto, and D. Capurro, “A survey on deep learning and explainability for automatic report generation from medical images,” *ACM Computing Surveys (CSUR)*, vol. 54, no. 10s, pp. 1–40, 2022.
- [2] Y. Zhou, F. Ringeval, and F. Portet, “A survey of evaluation methods of generated medical textual reports,” in *Proceedings of the 5th Clinical Natural Language Processing Workshop*, T. Naumann, A. Ben Abacha, S. Bethard, K. Roberts, and A. Rumshisky, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 447–459. [Online]. Available: <https://aclanthology.org/2023.clinicalnlp-1.48>
- [3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [4] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, “Generating radiology reports via memory-driven transformer,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Nov. 2020.
- [5] Z. Wang, H. Han, L. Wang, X. Li, and L. Zhou, “Automated radiographic report generation purely on transformer: A multicriteria supervised approach,” *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp. 2803–2813, 2022.
- [6] D. You, F. Liu, S. Ge, X. Xie, J. Zhang, and X. Wu, “Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation,” in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*. Springer, 2021, pp. 72–82.
- [7] Y. Wang, Z. Lin, and H. Dong, “Rethinking medical report generation: Disease revealing enhancement with knowledge graph,” *arXiv preprint arXiv:2307.12526*, 2023.
- [8] H. Lu, Q. Zhou, N. Fei, Z. Lu, M. Ding, J. Wen, C. Du, X. Zhao, H. Sun, H. He *et al.*, “Multimodal foundation models are better simulators of the human brain,” *arXiv preprint arXiv:2208.08263*, 2022.
- [9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” 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)*, 2019, pp. 4171–4186.
- [10] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked autoencoders are scalable vision learners,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.
- [11] X. Geng, H. Liu, L. Lee, D. Schuurmans, S. Levine, and P. Abbeel, “Multimodal masked autoencoders learn transferable representations,” in *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*.
- [12] G. Li, H. Zheng, D. Liu, C. Wang, B. Su, and C. Zheng, “Semmac: Semantic-guided masking for learning masked autoencoders,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 14 290–14 302, 2022.
- [13] A. J. Cawsey, B. L. Webber, and R. B. Jones, “Natural language generation in health care,” *Journal of the American Medical Informatics Association*, vol. 4, no. 6, pp. 473–482, 1997.
- [14] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.
- [15] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara, “From show to tell: A survey on image captioning,” *arXiv preprint arXiv:2107.06912*, 2021.

- [16] Z. Wang, M. Tang, L. Wang, X. Li, and L. Zhou, "A medical semantic-assisted transformer for radiographic report generation," in *Medical Image Computing and Computer Assisted Intervention-MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part III*. Springer, 2022, pp. 655–664.
- [17] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
- [18] Z. Chen, Y. Shen, Y. Song, and X. Wan, "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)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 5904–5914. [Online]. Available: <https://aclanthology.org/2021.acl-long.459>
- [19] J. Wang, A. Bhalerao, and Y. He, "Cross-modal prototype driven network for radiology report generation," in *European Conference on Computer Vision*. Springer, 2022, pp. 563–579.
- [20] G. Tiancheng, L. Dongnan, L. Zhiyuan, and C. Weidong, "Complex organ mask guided radiology report generation," *arXiv preprint arXiv:2311.02329*, 2023.
- [21] T. Tanida, P. Müller, G. Kaissis, and D. Rueckert, "Interactive and explainable region-guided radiology report generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7433–7442.
- [22] J. Zhang, X. Shen, S. Wan, S. K. Goudos, J. Wu, M. Cheng, and W. Zhang, "A novel deep learning model for medical report generation by inter-intra information calibration," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 10, pp. 5110–5121, 2023.
- [23] J. Xiao, Y. Bai, A. Yuille, and Z. Zhou, "Delving into masked autoencoders for multi-label thorax disease classification," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 3588–3600.
- [24] Z. Chen, Y. Du, J. Hu, Y. Liu, G. Li, X. Wan, and T.-H. Chang, "Multi-modal masked autoencoders for medical vision-and-language pre-training," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 679–689.
- [25] B. Huang, Z. Zhao, G. Zhang, Y. Qiao, and L. Wang, "Mgmae: Motion guided masking for video masked autoencoding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13 493–13 504.
- [26] Y. Xiong, B. Varadarajan, L. Wu, X. Xiang, F. Xiao, C. Zhu, X. Dai, D. Wang, F. Sun, F. Iandola *et al.*, "Efficientsam: Leveraged masked image pretraining for efficient segment anything," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 111–16 121.
- [27] J. H. Moon, H. Lee, W. Shin, Y.-H. Kim, and E. Choi, "Multi-modal understanding and generation for medical images and text via vision-language pre-training," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 12, pp. 6070–6080, 2022.
- [28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [29] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, and M. Lungren, "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)*, 2020, pp. 1500–1519.
- [30] H. Taud and J. Mas, "Multilayer perceptron (mlp)," *Geomatic approaches for modeling land change scenarios*, pp. 451–455, 2018.
- [31] D. Demner-Fushman, M. Kohli, M. Rosenman, S. Shooshan, L. Rodriguez, S. Antani, G. Thoma, and C. McDonald, "Preparing a collection of radiology examinations for distribution and retrieval," *Journal of the American Medical Informatics Association : JAMIA*, vol. 23, 07 2015.
- [32] A. Johnson, T. Pollard, S. Berkowitz, N. Greenbaum, M. Lungren, C.-y. Deng, R. Mark, and S. Horng, "Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific Data*, vol. 6, p. 317, 12 2019.
- [33] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02. USA: Association for Computational Linguistics, 2002, p. 311–318. [Online]. Available: <https://doi.org/10.3115/10733083.1073135>
- [34] A. Lavie and A. Agarwal, "Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments," in *Proceedings of the Second Workshop on Statistical Machine Translation*, ser. StatMT '07. USA: Association for Computational Linguistics, 2007, p. 228–231.
- [35] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: <https://aclanthology.org/W04-1013>
- [36] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.
- [37] W. Hou, K. Xu, Y. Cheng, W. Li, and J. Liu, "Organ: Observation-guided radiology report generation via tree reasoning," *arXiv preprint arXiv:2306.06466*, 2023.
- [38] Z. Huang, X. Zhang, and S. Zhang, "Kiut: Knowledge-injected u-transformer for radiology report generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 809–19 818.
- [39] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.
- [40] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.
- [41] F. Liu, S. Ge, and X. Wu, "Competence-based multimodal curriculum learning for medical 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)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 3001–3012. [Online]. Available: <https://aclanthology.org/2021.acl-long.234>
- [42] Y. Yang, J. Yu, J. Zhang, W. Han, H. Jiang, and Q. Huang, "Joint embedding of deep visual and semantic features for medical image report generation," *IEEE Transactions on Multimedia*, vol. 25, pp. 167–178, 2023.
- [43] S. Yang, X. Wu, S. Ge, Z. Zheng, S. K. Zhou, and L. Xiao, "Radiology report generation with a learned knowledge base and multi-modal alignment," *Medical Image Analysis*, vol. 86, p. 102798, 2023.
- [44] K. Zhang, H. Jiang, J. Zhang, Q. Huang, J. Fan, J. Yu, and W. Han, "Semi-supervised medical report generation via graph-guided hybrid feature consistency," *IEEE Transactions on Multimedia*, vol. 26, pp. 904–915, 2024.
- [45] Z. Liu, Z. Zhu, S. Zheng, Y. Zhao, K. He, and Y. Zhao, "From observation to concept: A flexible multi-view paradigm for medical report generation," *IEEE Transactions on Multimedia*, vol. 26, pp. 5987–5995, 2024.
- [46] R. Liu, M. Li, S. Zhao, L. Chen, X. Chang, and L. Yao, "In-context learning for zero-shot medical report generation," in *Proceedings of the 32nd ACM International Conference on Multimedia*, ser. MM '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 8721–8730. [Online]. Available: <https://doi.org/10.1145/3664647.3680760>
- [47] Z. Chen, Y. Li, Z. Wang, P. Gao, J. Barthelemy, L. Zhou, and L. Wang, "Enhancing radiology report generation via multi-phased supervision," *IEEE Transactions on Medical Imaging*, pp. 1–1, 2025.
- [48] M. Zhao, J. Liu, H. Shen, B. Yan, M. Pei, and Y. Wang, "Integrating clinical knowledge and imaging for medical report generation," *Pattern Recognition Letters*, vol. 195, pp. 59–65, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0167865525001783>