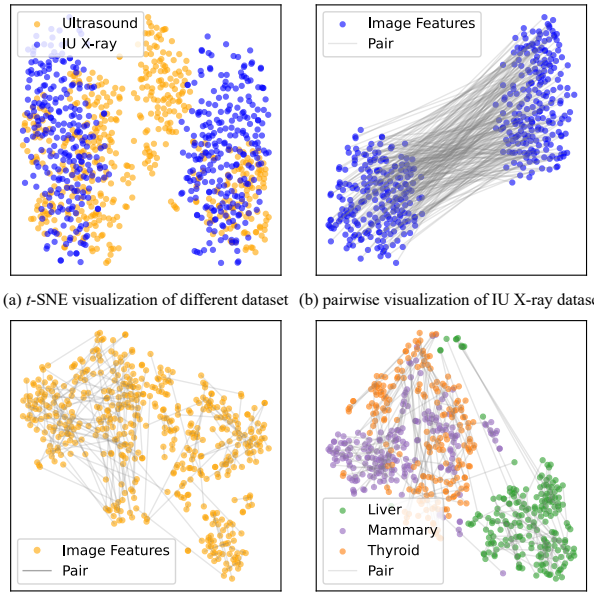
CURDNet: Contrastive Ultrasound Report Generation with Diversity-Aware Learning Network

Anonymous submission

Abstract

Ultrasound Report Generation (URG) aims to automatically produce diagnostic reports from ultrasound images, significantly reducing the workload of sonographers. However, URG remains underexplored due to the scarcity of high-quality datasets and the inherent diversity of ultrasound data. Unlike other radiology modalities such as X-rays, ultrasound imaging spans multiple organs and varies significantly with operator technique, leading to highly diverse visual appearances and reporting styles. This diversity complicates the design of generalizable report generation models, and overlooking it can negatively impact model performance. In this work, we propose a **C**ontrastive **U**ltrasound **R**eport-generation with **D**iversity-aware learning **Net**work, termed **CURDNet**, which explicitly accounts for the diverse characteristics of ultrasound data. Specifically: (1) we introduce **EchoDice**, a diversity-aware sampler that assembles training batches with high intra-batch variation to mimic cross-organ learning behavior and improve generalization; (2) we jointly train **ReportMatcher**, a contrastive module that distinguishes matched from mismatched report-image pairs via self-supervision, and **ReportGenerator**, which produces textual reports from ultrasound images; and (3) we propose **ReportJudger**, a large-language-model-based scoring module that evaluates the relevance of retrieved reports. Experiments on a representative URG benchmark demonstrate that CURDNet outperforms existing methods from both ultrasound-specific and general radiology domains. We hope CURDNet serves as a strong and extensible baseline for future ultrasound report generation research.



(a) *t*-SNE visualization of different dataset (b) pairwise visualization of IU X-ray dataset

(c) pairwise visualization of ultrasound dataset (d) organ distribution of ultrasound dataset

Figure 1: Comparison between IU X-ray and ultrasound report dataset. (a) shows image distribution of the two modality; (b) shows image pairs of IU X-ray; (c) shows image pairs of ultrasound report dataset; (d) shows ultrasound image paris by organs.

1 Introduction

Automatic report generation from medical images has gained increasing attention in recent years, driven by the promise of reducing clinicians' workload on both writing reports and teaching appearentees. Among existing studies, radiology report generation (RRG) has seen notable progress (Chen et al. 2020; Zhang et al. 2020; Liu et al. 2021; Li et al. 2023; Jin et al. 2024; Park et al. 2025), supported by large-scale radiology report datasets IU X-ray (Demner-Fushman et al. 2016), MIMIC-CXR (Johnson et al. 2019) and CheXpert (Irvin et al. 2019), which provided well-structured anotation by clinicians. In contrast, ultrasound report generation (URG), despite the ubiquity of ultrasound imaging in clinical practice, seen less activities due to the rareness of ultrasound report datasets (Li et al. 2025).

Ultrasound imaging, widely used in clinical settings for its low cost, safety, and real-time capability, differs fundamentally from other radiological modalities like X-ray and CT. It covers a broad range of organs and is highly operator-dependent, leading to image acquisition variability and diverse reporting styles. For instance, reports describing thyroid findings can differ significantly from those describing breast findings. As such, ultrasound report generation may also be more difficult for models compared with radiology case, as current radiology report generation datasets focus mainly on chest, which is in a fix style and describe certain objects. We further illustrate this in Figure 1. Figure 1a

gives an overview of these two different dataset, where IU X-ray dataset seems more consistent; Figure 1b,c outline the paired images of same patient for X-ray and ultrasound, respectively, indicating that X-rays are captured in standardized views (e.g., PA and lateral) and are more uniform. In contrast, ultrasound images lack such consistency, even within the same organ. As shown in Figure 1d, where different organs are highlighted in color, no clear pattern emerges. The disordered gray connecting lines reflect the high variability of ultrasound in both image composition and reporting style.

Due to the diversity of ultrasound, clinicians often need to repeatedly study different organs and cases in practice. Inspired by this, we raise the question: can we mimic the sonographer's learning behavior to develop an ultrasound report generation system that inherently handles diversity and distinguishes different cases? To this end, we introduce CURDNet, a novel framework that formulates ultrasound report generation as a mutual-task learning problem, jointly optimizing report generation and report-image retrieval. CURDNet comprises four key components: **EchoDice**, a diversity-aware sampling strategy to maximize the diversity of data loading; **ReportMatcher**, a contrastive learning module that try to distinguish different report pairs; **ReportGenerator**, trained for ultrasound report generation; and **ReportJudger**, in order to quantitatively assess the model's discriminative capability, which is a text-based scoring module leveraging large language method.

Our approach is straightforward yet quite effective, promoting better generalization and cross-organ understanding without requiring additional annotation. Experiments on a public ultrasound report generation benchmark demonstrate that CURDNet achieves competitive performance compared to state-of-the-art baselines originally designed for radiology tasks. These results suggest CURDNet provides a strong baseline and a practical foundation for future research in ultrasound report generation.

To summarize, our main contributions are as followed:

1. We introduce **EchoDice**, a diversity-aware sampling strategy aiming to maximize diversity during training.
2. We propose **ReportMatcher** and **ReportGenerator**, jointly mutual-task learning network that jointly trains a report generator and a contrastive retrieval module, effectively learning contrastive features and generate more accurate reports.
3. We introduce **ReportJuder**, a text-based scoring module that leverages large language models in order to quantitatively assess the discriminative quality of generated reports.

2 Related Work

2.1 Radiology Report Generation

Radiology Report Generation (RRG) has gained increasing attention with the availability of large-scale datasets such as IU X-ray (Demner-Fushman et al. 2016), MIMIC-CXR (Johnson et al. 2019), and CheXpert (Irvin et al. 2019). These datasets, together with clinical evaluation metrics, have driven the development of automated chest X-ray report generation systems. Most recent RRG methods adopted Transformer-based architectures inspired by image captioning (Vinyals et al. 2015), enabling the modeling of long-range dependencies between visual and textual information.

Representative approaches include R2Gen (Chen et al. 2020), which augmented Transformers with a memory module to preserve key contextual information, and PPKED (Liu et al. 2021), which predicted disease-related topics before generating descriptive sentences. Subsequent works such as RGKE (Yang et al. 2022) incorporated medical knowledge graphs, while RGRG (Tanida et al. 2023) introduced region-guided mechanisms for anatomical grounding. More recent models like RG-MemoryAlign (Shen et al. 2024) and MLRG (Liu et al. 2025) focus on aligning visual-textual embeddings and leveraging multi-view longitudinal imaging to improve temporal consistency. However, most of these methods are designed for chest X-ray datasets with standardized acquisition protocols, making them less directly applicable to the operator-dependent and heterogeneous nature of ultrasound imaging. However, these methods are tailored to the relatively standardized characteristics of chest X-ray imaging, and lack consideration for the complex and diverse nature of ultrasound data.

2.2 Ultrasound Report Generation

Ultrasound Report Generation (URG) remains comparatively underexplored, primarily due to the scarcity of large-scale annotated datasets and the unique characteristics of ultrasound imaging. Unlike X-ray or CT, ultrasound examinations are highly operator-dependent, vary in imaging angles, and are performed across multiple anatomical regions (e.g., liver, thyroid, breast), each with different diagnostic focuses and reporting conventions. These factors result in greater variability in image appearance and textual descriptions, complicating the design of unified modeling frameworks.

Despite these challenges, a number of studies have made early progress. AMAnet (Yang et al. 2021a) proposed a multi-label classification network to extract local semantic cues for guiding report generation. Deng (Deng et al. 2024) further combined a Transformer-based language model with a memory module and a hierarchical semantic structure to produce more detailed descriptions. Most recently, KMVE (Li et al. 2025) introduced knowledge-aware visual feature enhancement, achieving state-of-the-art performance across multiple organ-specific subsets of a newly released ultrasound report dataset. These efforts highlight the need for modeling strategies that explicitly account for the diversity and complexity inherent to ultrasonography. Compared with these approaches, our method places greater emphasis on the diverse characteristics of ultrasound imaging and aims to mitigate the performance degradation caused by such variability during training.

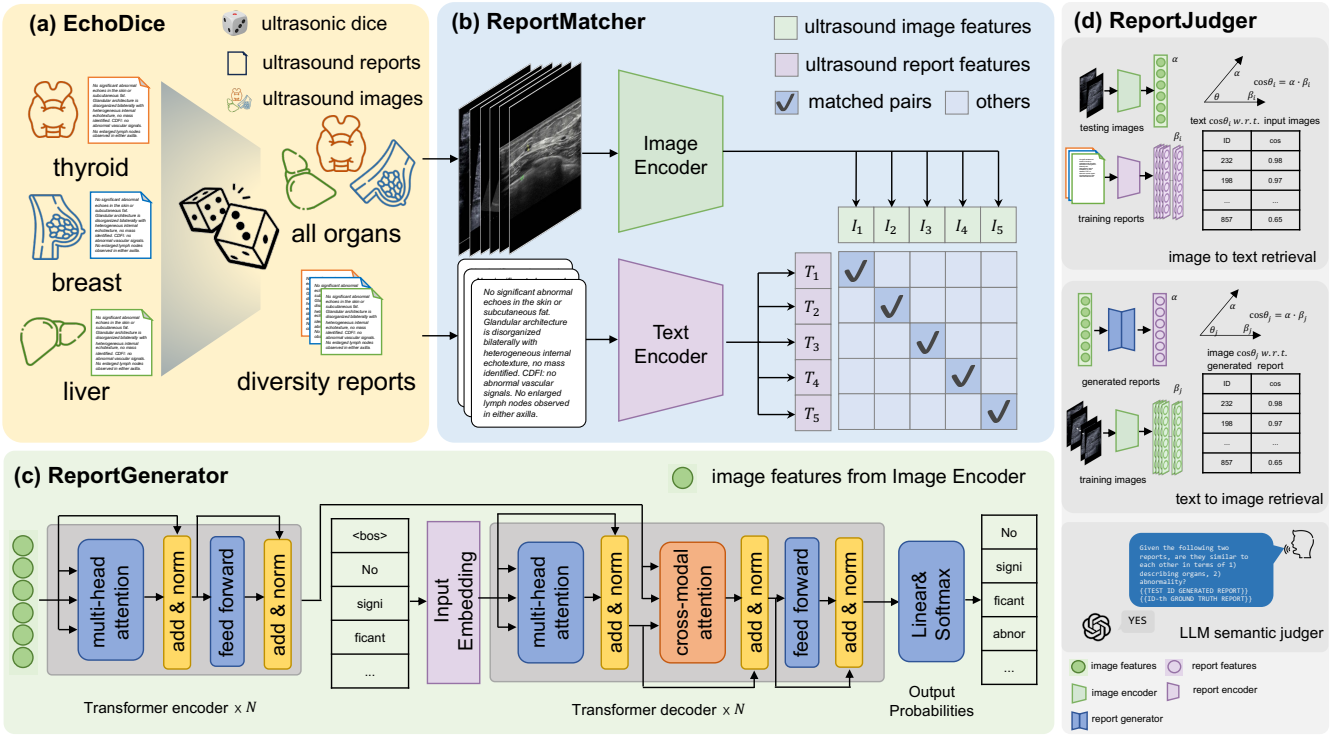


Figure 2: Architecture of the proposed method. (a) EchoDice is a "dice" that tries to filter data as much diversity as possible in each batch; (b) ReportMatcher consists of an image encoder and text encoder designed to distinguish different pairs of training samples; (c) ReportGenerator. The core part of our model for ultrasound report generation; (d) ReportJudger. An text-based LLM-aided tool for evaluating how well a matcher does.

3 Methodology

Overview

As shown in Figure 2, to tackle the challenges posed by the diversity of ultrasound data, we propose **EchoDice** to enable the model to learn from the most diverse data simultaneously. Furthermore, we leverage the joint training of **ReportMatcher** and **ReportGenerator** to enhance the model's learning capability in generating reports while distinguishing between different organ reports. Additionally, we introduce **ReportJudger**, a large-model-based pure text scorer, to evaluate the effectiveness of ReportMatcher.

3.1 EchoDice

In real-world ultrasound datasets, the distribution of organ categories is often highly imbalance — some organs are frequently imaged and richly annotated, while others are rare and sparsely represented. Training model accordingly can bias report generation models, limiting their generalization to underrepresented organs. To address this, we propose **EchoDice**, a diversity-aware sampling strategy designed to expose the model to a broader spectrum of semantic patterns during training. The name "Dice" reflects the stochastic, roll-like behavior of our sampler in assembling organ-diverse batches — akin to rolling a die where each number would have same probability to appear.

Given a dataset $\mathcal{D}$ containing n organ categories

$$\mathcal{D} = \{O_1, O_2, \ldots, O_n\},$$

where each category O_i contains N_i samples, we identify the most frequent class $O_{\max}$ with $N_{\max}$ samples and the rarest class $O_{\min}$ with $N_{\min}$ samples. Based on this, EchoDice supports two complementary strategies: up-sampling and down-sampling, both of which aim to ensure class-level balance and semantic diversity.

To amplify exposure to rare organ categories, we replicate instances from underrepresented classes so that they appear more frequently in training. Let the batch size be B and the number of categories per batch be k (typically $k \leq n$). We enforce B to be divisible by k to assign $\frac{B}{k}$ samples per category. In each iteration, we roll the "EchoDice" to uniformly sample k distinct categories and draw $\frac{B}{k}$ samples (with replacement if needed) from each. This up-sampling manner ensures every batch is semantically diverse and gives balanced attention to both common and rare organs.

Alternatively, a similar down-sampling can be applied to prevent dominant categories from overwhelming training by limiting each category to $N_{\min}$ instances. For categories where $N_i > N_{\min}$, we can select $N_{\min}$ samples without replacement. But we find that down-sampling behavior does not help improve the performance hence is not adopted.

During training, EchoDice dynamically assembles batches that are both balanced and semantically diverse.

However, this strategy is only applied during training. At inference time, the model operates on the natural data distribution without any sampling adjustments.

By focusing samely attention to each organs, EchoDice encourages the model to generalize better across organs, enhancing its capability to generate accurate, diverse, and robust medical reports.

3.2 ReportMatcher

To mimic how clinicians compare similar cases, we introduce **ReportMatcher**, a dual-encoder module aligning ultrasound images and reports in a shared embedding space.

Each training sample consists of an image pair $(\text{Img}_1, \text{Img}_2)$ and a report $T = [w_1, \cdots, w_L]$. Using a backbone F, image features are extracted as

$$[I_1, I_2] = [F(\text{Img}_1), F(\text{Img}_2)], \quad I_1, I_2 \in \mathbb{R}^{1024} \tag{1}$$

and aggregated via average pooling:

$$v = \frac{1}{2}(I_1 + I_2) \in \mathbb{R}^{1024}. \tag{2}$$

For text, a Sentence-BERT encoder generates hidden states

$$h = BERT(T) \in \mathbb{R}^{L \times d}, \tag{3}$$

from which the [CLS] token is taken as the sentence feature

$$t_{[CLS]} = h_{[CLS]} \in \mathbb{R}^d, \tag{4}$$

then projected into the image space:

$$t = proj(t_{[CLS]}) \in \mathbb{R}^{1024}. \tag{5}$$

A batch of B image-text pairs $\{(v_i, t_i)\}$ forms a similarity matrix $S_{i,j} = v_i^\top t_j$. We apply a symmetric contrastive loss:

$$\mathcal{L}_{i2t} = -\frac{1}{B}\sum_i \log \frac{\exp(S_{i,i}/\tau)}{\sum_j \exp(S_{i,j}/\tau)}, \tag{6}$$

$$\mathcal{L}_{t2i} = -\frac{1}{B}\sum_i \log \frac{\exp(S_{i,i}/\tau)}{\sum_j \exp(S_{j,i}/\tau)}, \tag{7}$$

$$\mathcal{L}_{\text{matcher}} = \frac{1}{2}(\mathcal{L}_{i2t} + \mathcal{L}_{t2i}), \tag{8}$$

where τ is a learnable temperature. This encourages correct image-text matches and separates mismatches in the embedding space.

3.3 ReportGenerator

The **ReportGenerator** generates ultrasound reports conditioned on visual input using a lightweight transformer-based encoder-decoder design with three transformer layers.

We reuse the visual backbone F from ReportMatcher and take its intermediate features $F[:-1]$ to form richer visual tokens. For two input images $(\text{Img}_1, \text{Img}_2)$, we compute the average embedding:

$$v = \frac{1}{2}(F[:-1](\text{Img}_1) + F[:-1](\text{Img}_2)) \in \mathbb{R}^{M \times d},$$

where M is the number of visual tokens. A transformer encoder is then applied to capture intra-visual dependencies:

$$\tilde{v} = \text{TransformerEncoder}(v).$$

For text decoder, it is a standard transformer decoder. Given partial report tokens $T = [w_1, \ldots, w_t]$, it first applies masked self-attention over text, then cross-attends to the encoded visual tokens $\tilde{v}$ using multi-head attention:

$$\text{Attn}(Q_t, K_v, V_v) = \text{softmax}\left(\frac{Q_t K_v^\top}{\sqrt{d}}\right) V_v.$$

The generator is trained with teacher-forcing and a cross-entropy loss:

$$\mathcal{L}_{\text{gen}} = -\sum_{t=1}^{L} \log P(w_t \mid w_{<t}, \tilde{v}),$$

where L is the length of the report. This setup encourages the model to align visual content with fluent text generation.

3.4 ReportJudger

To evaluate the semantic matching ability of the trained image and text encoders in ReportMatcher without relying on dense annotations, we propose a blind scoring scheme using a large language model (LLM). Specifically, we retrieve the top-K and bottom-K candidates by calculating cross-modality cosine similarity of image features and report features, and present each query-candidate pair individually to the LLM without revealing whether it belongs to the top or bottom set. The LLM is instructed to judge the semantic relevance between the query and each candidate description in a binary fashion (`relevant` or `irrelevant`). We provide more details regarding the prompt setting in supplementary materials.

Let q_i be a query (image or text), and let $\mathcal{R}_i = \{r_i^1, \ldots, r_i^{2K}\}$ denote the retrieved candidates, consisting of K top-ranked results and K bottom-ranked results. For each pair (q_i, r_i^j), the LLM outputs a binary label:

$$s_i^j = \text{LLM}(q_i, r_i^j) \in \{0, 1\},$$

where $s_i^j = 1$ indicates the LLM judges the pair to be semantically relevant.

We define the final retrieval consistency score over N queries as:

$$\text{LLM-T@K} = \frac{1}{N \cdot K}\sum_{i=1}^{N}\sum_{j=1}^{K} s_i^j \tag{9}$$

$$\text{LLM-B@K} = \frac{1}{N \cdot K}\sum_{i=1}^{N}\sum_{j=K+1}^{2K} s_i^j \tag{10}$$

where LLM-T@K represents the proportion of top-K results judged relevant by the LLM, and LLM-B@K denotes the proportion for the bottom-K results. A high LLM-T@K and low LLM-B@K indicate that the encoder retrieves semantically aligned content. As a scalar metric reflecting the overall separation between top and bottom retrieval sets. Intuitively, LLM-T@K close to 1 and LLM-B@K close to 0 suggest that the encoder retrieves semantically meaningful content at the top ranks while pushing irrelevant samples toward the bottom.

Split	Method	Publication	NLG METRICS ↑						CE METRICS ↑			
			BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	Accuracy	Precision	Recall	F1 Score
Mammary	CNN-RNN (Vinyals et al. 2015)	CVPR	0.114	0.093	0.078	0.067	0.221	0.185	0.000	0.496	0.498	0.487
	TriNet (Yang et al. 2021b)	TMM	0.693	0.594	0.533	0.478	0.439	0.742	0.351	0.816	0.697	0.727
	R2Gen (Chen et al. 2020)	EMNLP	0.663	0.611	0.572	0.541	0.411	0.685	0.494	0.800	0.761	0.776
	Transformer (Vaswani et al. 2017)	NIPS	0.699	0.653	0.619	0.590	0.437	0.757	0.461	0.827	0.671	0.702
	DeltaNet (Wu et al. 2022)	ACL	0.716	0.665	0.638	0.608	**0.517**	0.758	0.573	0.819	0.819	0.818
	R2GenRL (Qin and Song 2022)	ACL	0.672	0.595	0.531	0.479	0.500	0.651	0.424	0.793	0.754	0.771
	SGF (Li et al. 2025)	TMI	0.761	0.710	**0.672**	**0.640**	0.468	**0.758**	**0.586**	0.815	0.831	0.822
	Ours	-	**0.763**	**0.711**	0.670	0.637	0.470	0.755	0.547	**0.905**	**0.906**	**0.905**
Thyroid	CNN-RNN (Vinyals et al. 2015)	CVPR	0.131	0.105	0.086	0.069	0.069	0.207	0.000	0.448	0.348	0.382
	TriNet (Yang et al. 2021b)	TMM	0.645	0.510	0.421	0.345	0.409	0.678	0.268	0.845	0.769	0.803
	R2Gen (Chen et al. 2020)	EMNLP	0.578	0.532	0.492	0.457	0.369	0.664	0.404	0.810	0.768	0.779
	Transformer (Vaswani et al. 2017)	NIPS	0.709	0.642	0.585	0.538	0.425	0.701	0.260	0.717	0.732	0.724
	DeltaNet (Wu et al. 2022)	ACL	0.610	0.559	0.515	0.579	0.443	0.685	0.363	0.837	0.784	0.795
	R2GenRL (Qin and Song 2022)	ACL	0.616	0.595	0.464	0.414	**0.470**	0.599	0.434	0.834	0.819	0.826
	SGF (Li et al. 2025)	TMI	0.729	0.666	0.613	**0.568**	0.439	0.723	**0.524**	0.838	0.850	0.841
	Ours	-	**0.733**	**0.670**	**0.615**	**0.568**	0.440	**0.726**	0.514	**0.912**	**0.924**	**0.918**
Liver	CNN-RNN (Vinyals et al. 2015)	CVPR	0.049	0.026	0.011	0.000	0.119	0.102	0.000	0.181	0.068	0.070
	TriNet (Yang et al. 2021b)	TMM	0.868	0.821	0.785	0.750	0.531	0.861	0.039	0.898	0.809	0.814
	R2Gen (Chen et al. 2020)	EMNLP	0.866	0.842	0.822	0.805	0.537	0.869	0.530	0.875	0.880	0.870
	Transformer (Vaswani et al. 2017)	NIPS	0.855	0.832	0.815	0.800	0.524	0.873	0.444	0.749	0.785	0.765
	DeltaNet (Wu et al. 2022)	ACL	0.873	0.846	0.825	0.808	0.593	0.862	0.568	0.900	0.878	0.874
	R2GenRL (Qin and Song 2022)	ACL	0.853	0.818	0.791	0.769	0.575	0.842	0.466	0.885	0.875	0.879
	SGF (Li et al. 2025)	TMI	0.872	0.848	**0.828**	**0.813**	0.539	**0.875**	0.541	0.879	0.894	0.883
	Ours	-	**0.879**	**0.851**	**0.828**	0.810	**0.544**	0.874	**0.579**	**0.901**	**0.899**	**0.900**

Table 1: Comparison of ultrasound report generation in terms of NLG and clinical efficacy metrics among different state-of-the-art methods. **Best** results are in bold. Second best results are underlined.

4 Experiments

4.1 Settings

Dataset. Due to the scarcity of publicly available ultrasound image-report datasets, we evaluate our method on a recently released dataset by (Li et al. 2025), which does not contain any personally identifiable health information. The dataset covers three different types of organs, including the breast, thyroid, and liver. Specifically, the breast subset includes 3,521 patients, the thyroid subset contains 2,474 patients, and the liver subset comprises 1,395 patients. Each sample consists of a pair of ultrasound images (`Image1`, `Image2`) and a corresponding free-text finding. To ensure fair comparison, we adopt the official train/validation/test split provided in the dataset, with a ratio of 7:1:2 for each organ.

Metrics. We evaluate report generation performance using standard Natural Language Generation (NLG) metrics, including BLEU-1 to BLEU-4 (Papineni et al. 2002), METEOR (Denkowski and Lavie 2011), and ROUGE-L (Lin 2004). We also report the Clinical Efficacy (CE) following steps described by (Li et al. 2025). We use ReportJudger to evaluate the cross-modality retrieval performance. For all but LLM-B@K metrics, a higher value indicates better performance.

Training Settings. We adopt the pretrained CNN backbone ResNet101 (He et al. 2016) for both ReportMatcher and ReportGenerator; we use `bert-base-chinese` (Devlin et al. 2019) for sentence embedding. The network is trained under 100 epochs, where early stopping would apply if no improvement observed during 15 consecutive epochs. More settings can be found in *Supplementary Material*.

4.2 Results and Analysis

Generation Performance. Table 1 summarizes the performance of our method compared to several state-of-the-art approaches on the ultrasound report generation task, evaluated across three organ-specific subsets: mammary, thyroid, and liver. The table reports results in terms of both natural language generation (NLG) metrics and clinical efficacy (CE) metrics. **Best** results are highlighted in bold, and second-best results are underlined.

Overall, our proposed ReportGenerator consistently demonstrates superior performance. It achieves the highest BLEU-1 and BLEU-2 scores across all three organ subsets, indicating its capability in accurately reproducing relevant n-gram patterns from reference reports. For BLEU-3, BLEU-4, and ROUGE-L, our method ranks either first or second in all organ splits, showcasing its strength in capturing longer-term dependencies and overall fluency. Moreover, our method also keeps competitive scores in the METEOR metric.

In terms of clinical efficacy, which reflects the practical utility of generated reports in a medical setting, our method achieves top performance on Precision, Recall, and F1 Score metrics across all three organs. This underscores the clinical relevance and correctness of the generated content. Although our model's accuracy is slightly lower than the highest-performing baseline in this metric (particularly SGF), it remains comparable and still demonstrates robust performance in practice.

Unlike SGF (Li et al. 2025), which trains separate models for each organ to optimize performance, our approach employs a unified model architecture enhanced by *EchoDice* and *ReportMatcher*, yet still outperforms or closely matches

Split	Setting	B-1	B-4	MTR	R-L
Mammary	TF	0.699	0.590	0.437	**0.757**
	w/ Mat.	0.733	0.598	0.455	0.736
	w/ Dice.	0.729	0.586	0.446	0.731
	Ours	**0.763**	**0.637**	**0.470**	0.755
Thyroid	TF	0.709	0.538	0.425	0.701
	w/ Mat.	0.728	0.560	0.435	0.723
	w/ Dice.	0.690	0.529	0.416	0.715
	Ours	**0.733**	**0.568**	**0.440**	**0.726**
Liver	TF	0.855	0.800	0.524	0.873
	w/ Mat.	**0.879**	0.809	**0.544**	0.866
	w/ Dice.	**0.879**	**0.814**	**0.544**	0.871
	Ours	**0.879**	0.810	**0.544**	**0.874**

Table 2: Ablation study for ReportGenerator on other components of the proposed method. "TF" denotes the Transformer baseline without any additional components. "w/ Mat." indicates the model with the ReportMatcher, while "w/ Dice." refers to the model with the EchoDice. **Best** results are highlighted in bold.

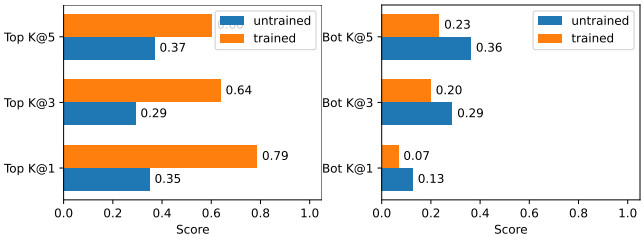
specialized models. This demonstrates the scalability and efficiency of our framework, achieving state-of-the-art results without the need for organ-specific training.

Ablation Study. To assess the individual contribution of each component in our proposed ReportGenerator, we conduct a thorough ablation study by incrementally incorporating the ReportMatcher and EchoDice modules, as presented in Table 2. The Transformer-only baseline (denoted as "TF") serves as a basic model for ReportGenerator. We then evaluate performance when adding only the ReportMatcher ("w/ Mat."), only EchoDice ("w/ Dice."), and both modules together ("Ours").

Across all organ subsets, integrating the **ReportMatcher** consistently improves performance over the baseline Transformer, particularly on BLEU-1 and METEOR. This demonstrates its effectiveness in refining the alignment between generated reports and clinically meaningful sentence structures by leveraging retrieved historical reports.

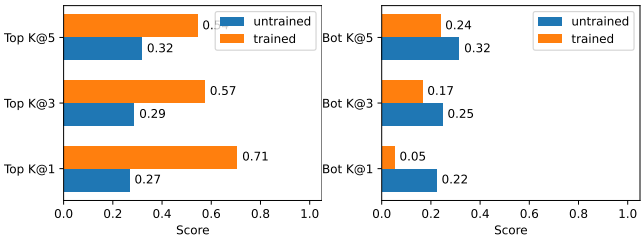
The **EchoDice** module, when used in isolation, shows mixed results. While it leads to performance gains in the Mammary subset -—- most notably improving BLEU-1 from 0.733 to 0.763 when used together with ReportMatcher —— it has a limited standalone effect for the Thyroid and Liver subsets. However, when combined with ReportMatcher, EchoDice still contributes to slight but consistent gains in multiple metrics, especially in longer n-gram scores such as BLEU-4 and ROUGE-L. This suggests that EchoDice complements semantic matching by injecting organ-specific prior knowledge at the token level.

For the Liver subset, our full model achieves identical or near-identical best scores across all metrics. This indicates that while the Transformer baseline already performs strongly for liver reports, the additional modules help further consolidate performance, particularly for METEOR and ROUGE-L.



(a) text to image, top K (b) text to image, bottom K

Figure 3: Text to imge retrieval performance. Both the top K and bottom K performance are significantly better than the baseline after our model training. This indicates that our model is able to retrieve semantically relevant images for a given text query.



(a) image to text, top K (b) image to text, bottom K

Figure 4: Image to text retrieval performance. Both the top K and bottom K performance are significantly better than the baseline after our model training. This indicates that our model is able to retrieve semantically relevant texts for a given image query.

In summary, both modules provide complementary benefits: ReportMatcher enhances textual alignment and factual consistency, while EchoDice contributes domain-specific structural priors. Their combined use results in the best overall performance across most metrics and organ types, validating the effectiveness of our CURDNet.

ReportMatcher Performance. To assess the effectiveness of the proposed ReportMatcher, we utilize ReportJudger to compare semantic and clinical relevance metrics before and after training. The evaluation results across three organs are illustrated in Figures 3 and 4.

In these figures, LLM-T@1 (also labeled as Top K@1) represents the proportion of pairs deemed semantically and clinically relevant by the large language model (LLM). For instance, a pair where one report describes the thyroid and the other describes the liver would be classified as `irrelevant`. Similarly, a report about a thyroid nodule and another about the general thyroid gland would also be considered `irrelevant`. Only reports describing the same clinical condition are labeled as `relevant`.

As shown in Figure 3a and Figure 4a, LLM-T@1 significantly increases after training, indicating that ReportMatcher effectively aligns cross-modal representations. For example, in the $i \rightarrow t$ direction, Top K@1 improves from 0.27 to 0.71, demonstrating the model's ability to bring se-

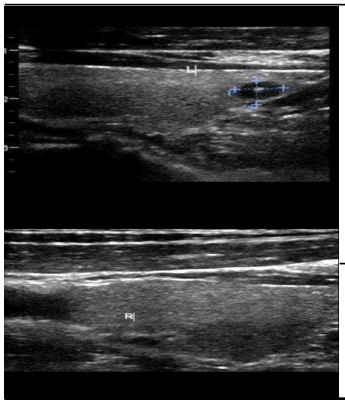
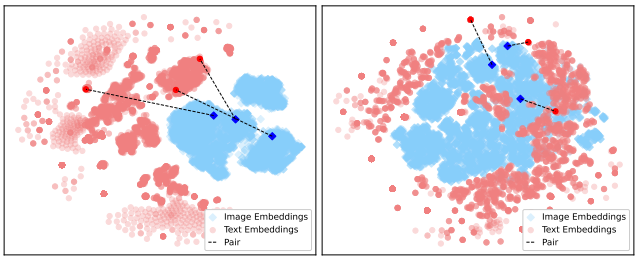
Image	Ground Truth	Vanilla TF	Ours
	The thyroid gland is normal in size and morphology. A cystic nodule is visualized at the lower pole of the left lobe, measuring approximately _2DS_, with well-defined margins and a regular shape. Color Doppler Flow Imaging (CDFI) shows no detectable blood flow signal. The remaining thyroid parenchyma demonstrates homogeneous echogenicity, and CDFI reveals no abnormal intraglandular blood flow signals. No significantly enlarged lymph nodes are detected in the bilateral cervical regions.	The thyroid gland is normal in size and morphology, with homogeneous echogenicity. No definite space-occupying lesion is identified. CDFI reveals no abnormal intraglandular blood flow signals. Multiple hypoechoic nodules are visualized in the bilateral cervical regions, the largest measuring approximately _2DS_ on the left and _2DS_ on the right, with well-defined margins and regular shapes. The "hilum sign" structure is observed, and CDFI demonstrates detectable blood flow signals.	The thyroid gland is normal in size and morphology. A cystic nodule is visualized in the left lobe, measuring approximately _2DS_, with well-defined margins and a regular shape. CDFI shows no detectable blood flow signal. The remaining thyroid parenchyma demonstrates homogeneous echogenicity, and CDFI reveals no abnormal intraglandular blood flow signals. No significantly enlarged lymph nodes are detected in the bilateral cervical regions.
	甲状腺大小形态如常，左叶下极可见一囊性结节，大小约_2DS_，边界清晰，形态规整，CDFI示未探及血流信号；余腺体回声均匀，CDFI示腺体内未见异常血流信号。双侧颈部未见明显肿大淋巴结。	甲状腺大小形态如常，腺体回声均匀，未见明确占位性病变，CDFI示腺体内未见异常血流信号。双侧颈部扫查可探及多个低回声结节，左侧大者约_2DS_，右侧大者约_2DS_，边界清晰，形态规整，可见"淋巴门"结构，CDFI示可探及血流信号。	甲状腺大小形态如常，左叶可见一囊性低回声结节，大小约_2DS_，边界清晰，形态规整，CDFI示未探及血流信号。余腺体回声均匀，CDFI示腺体内未见异常血流信号。双侧颈部未见明显肿大淋巴结。

Figure 5: Comparison of generated ultrasound reports. Text highlighted in green means that the words are both in generated text and ground truth. Text highlighted in yellow means the ground truth words are not generated. We translate the original Chinese text using ChatGPT for clarity.



(a) t-SNE before training (b) t-SNE after training

Figure 6: t-SNE visualization of the joint embedding space before and after training. The left figure shows the untrained model, where image and text embeddings are not well aligned. The right figure shows the trained model, where image and text embeddings are more closely aligned.

mantically matching image-report pairs closer together.

Moreover, in Figure 4b, the Bot K@1 (representing the worst-matching pairs) decreases from 0.22 before training to 0.05 after training. This suggests that the model also successfully pushes apart unrelated pairs, reducing the proportion of mismatched examples from 22% to just 5%.

To better understand cross-modal fusion, we conduct a t-SNE visualization shown in Figure 6. Before training (Figure 6a), image embeddings form three distinct clusters by organ category, with large distances to text embeddings, indicating poor alignment. After training (Figure 6b), embeddings become more compact and cross-modal distances significantly decrease. Three random image-report pairs highlight the reduced image-to-text distances. Nevertheless, as emphasized by "Mind The Gap" (Liang et al. 2022), a residual gap between image and text clusters remains, showing that perfect fusion is still challenging and an open problem.

Qualitative Study. Figure 5 shows a comparison of our ReportGenerator with Vanilla Transformer model. As there exists more green highlighted text in our model, it indicates that our method could generate more accurate reports than vanilla Transformer model. More case studies are presented in the *Supplementary Material*.

System Demonstration. To further showcase the functionality of our ultrasound report generation system, we built a webpage based on CURDNet that integrates report generation, similar case comparison, and archiving functions into a single platform. More details about the system demo and usage are available in the *Supplementary Material*.

6 Limitation and Discussion

Despite its promising performance in ultrasound report generation, our method has limitations. As EchoDice relies on organ labels, its generalizability may be restricted, which we address with an additional experiment in the *Supplementary Material*. Exploring different backbone architectures could further enhance understanding of the approach.

7 Conclusion

In this work, we propose a unified framework for ultrasound report generation and case retrieval, inspired by the way sonographers learn from diverse cases across different organs. Our method incorporates a diversity-aware sampling strategy to expose the model to heterogeneous cases during training, and a lightweight image-text contrastive learning objective to enhance the model's ability to understand and align similar image-text pairs. Experiments on a public ultrasound report dataset demonstrate that our method achieves superior performance compared with state-of-the-art baselines.

References

Chen, Z.; Song, Y.; Chang, T.-H.; and Wan, X. 2020. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*.

Demner-Fushman, D.; Kohli, M. D.; Rosenman, M. B.; Shooshan, S. E.; Rodriguez, L.; Antani, S.; Thoma, G. R.;

and McDonald, C. J. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2): 304–310.

Deng, J.; Chen, D.; Zhang, C.; and Dong, Y. 2024. Generating lymphoma ultrasound image description with transformer model. *Computers in Biology and Medicine*, 174: 108409.

Denkowski, M.; and Lavie, A. 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In Callison-Burch, C.; Koehn, P.; Monz, C.; and Zaidan, O. F., eds., *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 85–91. Edinburgh, Scotland: Association for Computational Linguistics.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. 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)*, 4171–4186.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 590–597.

Jin, H.; Che, H.; Lin, Y.; and Chen, H. 2024. Promptmrg: Diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2607–2615.

Johnson, A. E.; Pollard, T. J.; Berkowitz, S. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.-y.; Mark, R. G.; and Horng, S. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1): 317.

Li, J.; Su, T.; Zhao, B.; Lv, F.; Wang, Q.; Navab, N.; Hu, Y.; and Jiang, Z. 2025. Ultrasound Report Generation With Cross-Modality Feature Alignment via Unsupervised Guidance. *IEEE Transactions on Medical Imaging*, 44(1): 19–30.

Li, M.; Lin, B.; Chen, Z.; Lin, H.; Liang, X.; and Chang, X. 2023. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3334–3343.

Liang, V. W.; Zhang, Y.; Kwon, Y.; Yeung, S.; and Zou, J. Y. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35: 17612–17625.

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

Liu, F.; Wu, X.; Ge, S.; Fan, W.; and Zou, Y. 2021. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13753–13762.

Liu, K.; Ma, Z.; Kang, X.; Li, Y.; Xie, K.; Jiao, Z.; and Miao, Q. 2025. Enhanced contrastive learning with multi-view longitudinal data for chest x-ray report generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 10348–10359.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Isabelle, P.; Charniak, E.; and Lin, D., eds., *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Park, S.; Heo, K.; Shin, D.; Son, Y.; Oh, J.-H.; and Kam, T.-E. 2025. DART: Disease-aware Image-Text Alignment and Self-correcting Re-alignment for Trustworthy Radiology Report Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3334–3343.

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

Shen, H.; Pei, M.; Liu, J.; and Tian, Z. 2024. Automatic radiology reports generation via memory alignment network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4776–4783.

Tanida, T.; Müller, P.; Kaissis, G.; and Rueckert, D. 2023. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7433–7442.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3156–3164.

Wu, X.; Yang, S.; Qiu, Z.; Ge, S.; Yan, Y.; Wu, X.; Zheng, Y.; Zhou, S. K.; and Xiao, L. 2022. DeltaNet: Conditional medical report generation for COVID-19 diagnosis. *arXiv preprint arXiv:2211.13229*.

Yang, S.; Niu, J.; Wu, J.; Wang, Y.; Liu, X.; and Li, Q. 2021a. Automatic ultrasound image report generation with adaptive multimodal attention mechanism. *Neurocomputing*, 427: 40–49.

Yang, S.; Wu, X.; Ge, S.; Zhou, S. K.; and Xiao, L. 2022. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical image analysis*, 80: 102510.

Yang, Y.; Yu, J.; Zhang, J.; Han, W.; Jiang, H.; and Huang, Q. 2021b. Joint embedding of deep visual and semantic features for medical image report generation. *IEEE Transactions on Multimedia*, 25: 167–178.

Zhang, Y.; Wang, X.; Xu, Z.; Yu, Q.; Yuille, A.; and Xu, D. 2020. When radiology report generation meets knowledge graph. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 12910–12917.

Reproducibility Checklist

Instructions for Authors:
This document outlines key aspects for assessing reproducibility. Please provide your input by editing this `.tex` file directly.

For each question (that applies), replace the "Type your response here" text with your answer.

Example: If a question appears as

```
\question{Proofs of all novel claims
are included} {(yes/partial/no)}
Type your response here
```

you would change it to:

```
\question{Proofs of all novel claims
are included} {(yes/partial/no)}
yes
```

Please make sure to:

- Replace ONLY the "Type your response here" text and nothing else.
- Use one of the options listed for that question (e.g., **yes**, **no**, **partial**, or **NA**).
- **Not** modify any other part of the `\question` command or any other lines in this document.

You can `\input` this .tex file right before `\end{document}` of your main file or compile it as a stand-alone document. Check the instructions on your conference's website to see if you will be asked to provide this checklist with your paper or separately.

1. General Paper Structure

1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) yes

1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) yes

1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) yes

2. Theoretical Contributions

2.1. Does this paper make theoretical contributions? (yes/no) no

If yes, please address the following points:

2.2. All assumptions and restrictions are stated clearly and formally (yes/partial/no) Type your response here

2.3. All novel claims are stated formally (e.g., in theorem statements) (yes/partial/no) Type your response here

2.4. Proofs of all novel claims are included (yes/partial/no) Type your response here

2.5. Proof sketches or intuitions are given for complex and/or novel results (yes/partial/no) Type your response here

2.6. Appropriate citations to theoretical tools used are given (yes/partial/no) Type your response here

2.7. All theoretical claims are demonstrated empirically to hold (yes/partial/no/NA) Type your response here

2.8. All experimental code used to eliminate or disprove claims is included (yes/no/NA) Type your response here

3. Dataset Usage

3.1. Does this paper rely on one or more datasets? (yes/no) yes

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) yes

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) NA

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) NA

3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) yes

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) yes

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing (yes/partial/no/NA) NA

4. Computational Experiments

4.1. Does this paper include computational experiments? (yes/no) yes

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of

the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) yes

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) yes

4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no) yes

4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no) yes

4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no) partial

4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA) yes

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) yes

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) yes

4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) yes

4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) no

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) no

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) yes