UniMRG: Refining Medical Semantic Understanding Across Modalities via LLM-Orchestrated Synergistic Evolution

Hongyan Xu[1,2], Arcot Sowmya[1], Ian Katz[3], Dadong Wang[2⋆]

[1]School of Computer Science and Engineering, University of New South Wales
[2]Data61, The Commonwealth Scientific and Industrial Research Organisation
[3]Southern Sun Pathology Pty Ltd
hongyan.xu@unsw.edu.au, Dadong.wang@csiro.au,
a.sowmya@unsw.edu.au,ian.katz@southernsun.com.au

Abstract. Current medical report generation (MRG) methods remain limited by cross-modal associations, particularly when handling complex medical terminology across different modalities. In this work, we propose the Universal Medical Report Generation (UniMRG) framework to enhance Vision-Language foundation models (VLFMs) through coordinated data augmentation and architecture optimization. Specifically, we introduce Universal Semantics-Synergistic Multimodal Augmentation to enhance model adaptability to diverse medical scenarios while preserving critical diagnostic features. We further design a Medical Content Learner to capture both fine-grained pathological variations and specialized diagnostic contexts for robust cross-modal alignment. To achieve robust medical understanding against real-world variations, we develop a Dynamic Synergistic Evolution strategy guided by Large Language Model (LLM) that enables joint optimization of augmentation policies and architectural configurations. To address the existing gap in public VL datasets for skin diseases, we release a large-scale Skin-Path dataset, consisting of 277,761 patches covering 10 distinct skin diseases. Extensive experiments on PatchGastric22, IU-Xray, and Skin-Path demonstrate that UniMRG achieves state-of-the-art performance, surpassing Clinical-BERT by 2.6% in BLEU-4 and 3.9% in Rouge-L on IU-Xray. The Skin-Path dataset is available at: `https://unimrg.github.io/Skin-Path/`.

Keywords: Medical Report Generation · Cross-Modal Alignment · Large Language Models (LLMs).

1 Introduction

Medical report generation (MRG) is a key component supporting medical image computing and computer-aided diagnosis. It aims to generate accurate and coherent text from medical images such as X-rays [10], surgical images [11], and pathology slides [24], thereby assisting clinicians in diagnosis and improving

⋆ Corresponding author.

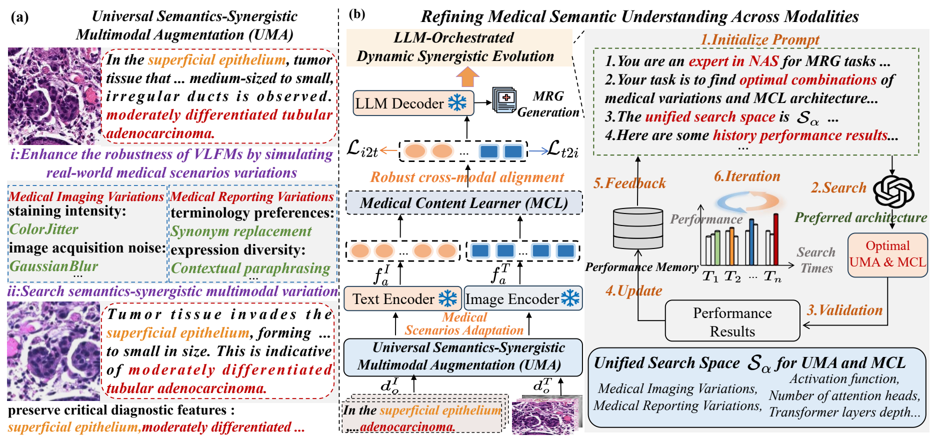


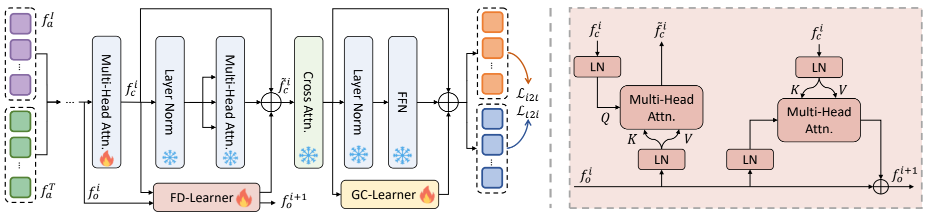
Fig. 1: Details of the proposed **UniMRG** method for medical report generation.

efficiency. Traditional methods [8,26,29] often exhibit limited generative capabilities, struggling to produce coherent long-form text. Although Transformer-based models [1,18] have improved long-range dependency modeling, they still face challenges in effectively aligning visual and textual information, particularly when handling complex medical images and specialized medical terminology [13].

Recent advances in Vision-Language Foundation Models (VLFMs) have enabled coherent, context-aware text generation [14,23,30]. While some works apply VLFMs to MRG [9,33], they struggle to maintain semantic consistency across modalities under medical scenario variations. Dependence on fixed augmentations [31,32,4] and static architectures [27,17,16] further limits adaptability. Key challenges include: ***i***) achieving accurate cross-modal alignment, especially in associating complex medical terminology with visual features; and ***ii***) designing robust modules that fully leverage VLFMs for cross-modal learning.

In this work, we propose the UniMRG framework, which incorporates the Universal Semantics-Synergistic Multimodal Augmentation (UMA) and the Medical Content Learner (MCL) modules to enhance VLFMs' understanding of medical imaging scenarios across diverse MRG tasks through data augmentation and architecture optimization. Furthermore, we design the Dynamic Synergistic Evolution strategy to jointly optimize augmentation policies and architectural configurations. The main contributions are as follows:

- We proposed the UniMRG framework from both data and structural perspectives, significantly enhancing the general VLFM for MRG tasks.
- We designed a Dynamic Synergistic Evolution method to explore the optimal model architecture and multi-modal augmentation strategy for MRG model.
- We introduced Skin-Path, which, to our best knowledge, is the first VL dataset for skin cancer, facilitating comprehensive evaluation of MRG tasks.
- Experiments on PatchGastric22, IU-Xray, and Skin-Path confirm UniMRG's effectiveness across medical domains (*e.g.,* pathology and chest X-rays).



(a) FD- and GC-Learner in the hierarchical Cross-modal Interaction Module (b) Multi-scale content interaction of FD-Learner in stage i

Fig. 2: The Medical Content Learner architecture for capturing variations and associations between augmented visual semantics and medical terminology.

Algorithm 1: LLM-Orchestrated Dynamic Synergistic Evolution

Input : LLM proxy $\mathcal{P}$, unified search space $\mathcal{S}_\alpha$, training dataset $\mathcal{D}_{\text{tr}}$, validation dataset $\mathcal{D}_{\text{val}}$, LLM search times $N_{\max}$, training epochs T.

Initialize $\mathcal{A}_{aug}$ and MCL configurations, evaluation metrics $Perf = \{\}$.
for N=0 *to* N_{max} **do**
 // Stage 1: Universal Semantics-Synergistic Multimodal Augmentation
 Generate optimal multimodal features (f_a^I, f_a^T) via $\mathcal{P}$ using Eq. 1
 // Stage 2: MCL Architecture Adaptation
 Update MCL architecture and extract cross-modal features using Eq. 6
 // Stage 3: Joint Training and Evaluation
 Train UniMRG for T epochs and evaluate to get $Perf_{\text{val}}$ using Eq. 4, 5
 // Stage 4: LLM-guided Synergistic Evolution
 Update augmentation and architecture configurations using Eq. 7, 8
end
Output: Optimal augmentation strategy $\mathcal{A}^*_{aug}$ and MCL architecture a^*.

2 Methodology

This section presents the UniMRG framework (Fig. 1), which includes: (a) Universal Semantics-Synergistic Multimodal Augmentation (Sec 2.1) to simulate real-world medical variations while preserving diagnostic features; and (b) Dynamic Synergistic Evolution (Sec 2.2), where an LLM proxy coordinates augmentation and architecture adaptation to enhance understanding of specialized medical terminology. We also introduce Skin-Path (Sec 2.3), the first VL dataset for skin cancer, supporting comprehensive MRG evaluation.

2.1 Universal Semantics-Synergistic Multimodal Augmentation

To enhance MRG capability using cross-modal medical data, we propose Universal Semantics-Synergistic Multimodal Augmentation (UMA), which aims to simulate diverse real-world medical variations while preserving key diagnostic features. UMA applies two strategies: $\mathcal{S}_I$ for images and $\mathcal{S}_T$ for text.

Let $\mathcal{S}_I$ and $\mathcal{S}_T$ represent image and text augmentation strategies (*e.g.*, RandomResizedCrop, ColorJitter for images; Synonym Replacement, Back-Translation

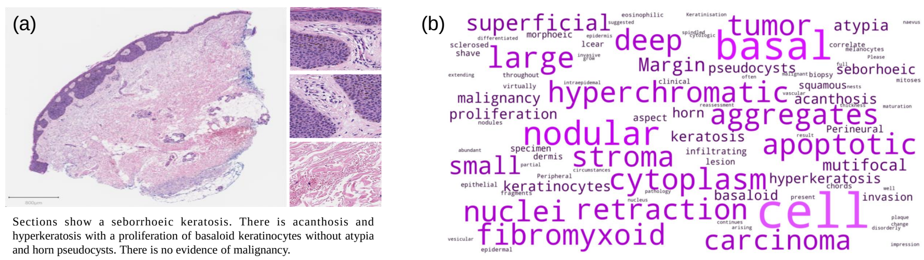


Fig. 3: (a) Example of WSI with medical report and patches. (b) word cloud.

for text). A task-aware LLM proxy $\mathcal{P}$, guided by prompts (*e.g.*, diagnostic focus), generates and refines augmentation pairs using validation feedback. Visual context is integrated via MCL-processed features. This process is defined as:

$$f_a^I, f_a^T = E_I(\mathcal{P}(\mathcal{S}_I, \mathcal{S}_T), d_o^I), E_T(\mathcal{P}(\mathcal{S}_I, \mathcal{S}_T), d_o^T), \tag{1}$$

where d_o^I, d_o^T represent the original medical image and report text, $\mathcal{P}$ denotes the LLM-guided proxy over both modalities, and E_I and E_T are image and text encoders, respectively, in the VLFM. The extracted features f_a^I, f_a^T aim to reflect real-world variations while preserving medical patterns, serving as the data foundation for robust cross-modal alignment in MRG tasks.

2.2 Dynamic Synergistic Evolution via LLM Proxy

Medical Content Learner (MCL). General VLFMs commonly lack the ability to associate and align medical terminology with visual semantics. To address this, we design a multi-level cross-modal content learner, comprising the FD (Fine-grained Details)-Learner and GC (Global Contexts)-Learner, as illustrated in Figure 2. While Figure 2 presents the high-level structure, the FD-Learner is followed by an external MHA and skip connection module that further refines the visual features. The interaction modeling process is formulated as:

$$\widetilde{f_c^i} = \text{FD-Learner}(f_o^i, f_c^i) + \text{LN}(\text{MHA}(f_o^i)) + f_o^i, \tag{2}$$

$$\widetilde{f_c^{i+1}} = \text{GC-Learner}(\text{CA}(\widetilde{f_c^i})) + \text{FFN}(\text{LN}(\text{CA}(\widetilde{f_c^i}))) + \text{CA}(\widetilde{f_c^i}). \tag{3}$$

where f_o^i and f_c^i are modality features from the i-th cross-modal interaction layer; $\widetilde{f_c^i}$ and $\widetilde{f}^{i+1}c$ are refined features capturing fine-grained pathology and diagnostic context. FD-Learner boosts sensitivity to visual-textual details; GC-Learner models global semantics via LoRA-based adaptation [6]. LN (LayerNorm) stabilizes the activations, MHA (Multi-Head Attention) captures intra-modal dependencies, and CA (cross-attention) aligns cross-modal features. Following [20], we use bidirectional contrastive losses ($\mathcal{L}i2t$, $\mathcal{L}_{t2i}$) for better alignment.

Dynamic Synergistic Evolution (DSE). In practical implementation, MRG models still face two key challenges: i) maintaining semantic consistency

Table 1: Comparative results on PatchGastric22 [24] and IU-Xray [2].

Dataset	Model	BL-1	BL-2	BL-3	BL-4	MTR	RG-L	C
PatchGastric22	EfficientNetB3 [22]	-	-	-	0.324	-	-	-
	VGG [21]	0.503	0.382	0.343	0.248	-	-	-
	ResNet [5]	0.503	0.382	0.343	0.248	-	-	-
	PVT [28]	0.503	0.382	0.343	0.248	-	-	-
	SWIN [16]	0.498	0.373	0.328	0.231	-	-	-
	ConvNeXt [17]	0.510	0.392	0.351	0.255	-	-	-
	UniMRG (Ours)	**0.541**	**0.465**	**0.408**	**0.368**	**0.306**	**0.529**	**2.670**
IU-Xray	ST [26]	0.216	0.124	0.087	0.066	-	0.306	0.277
	CoAtt [8]	0.455	0.288	0.205	0.154	-	0.369	0.277
	MRMA [29]	0.457	0.295	0.212	0.157	0.180	0.353	0.244
	HRGR [10]	0.438	0.298	0.208	0.151	-	0.322	0.343
	CMAS-RL [7]	0.464	0.301	0.210	0.154	-	0.362	0.275
	R2Gen [1]	0.470	0.304	0.219	0.165	0.187	0.371	-
	PPKED [15]	0.483	0.315	0.224	0.168	-	0.376	0.351
	CMN+MHAA [27]	0.503	0.328	0.232	0.172	0.212	0.395	-
	S3-NET [18]	0.499	0.334	0.246	0.172	0.206	0.401	-
	SILC [13]	0.472	0.321	0.234	0.175	0.192	0.379	0.368
	UniMRG (Ours)	**0.509**	**0.336**	**0.252**	**0.196**	**0.228**	**0.415**	**0.402**

between modalities after augmentation, and *ii*) adapting the MCL architecture to data distribution shifts. To address these, we introduce DSE, which uses an LLM as a task-aware controller to jointly optimize the UMA policy $\mathcal{A}aug$ and MCL configuration a (*e.g.*, module depth, attention heads) within a unified search space $\mathcal{S}\alpha$. This enables efficient data-model co-optimization with minimal cost, formulated as a neural architecture search (NAS) task:

$$\mathcal{W}^*(a) = \arg\min_{\mathcal{W}} \mathbb{E}\left[\mathcal{L}_{tr}(a, \mathcal{W}; \mathcal{A}_{aug}, \mathcal{D}_{tr})\right], \tag{4}$$

$$(a^*, \mathcal{A}^*_{aug}) = \underset{a\in\mathcal{S}_a, \mathcal{A}_{aug}\in\mathcal{S}_\alpha}{\arg\max} Perf_{\text{val}}\left(\mathcal{D}_{val}, \mathcal{A}_{aug}; a, \mathcal{W}^*, \mathcal{S}_\alpha\right). \tag{5}$$

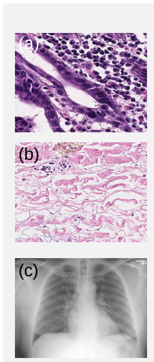
Here, $\mathcal{W}^*$ represents the weights of the optimal architecture a^*. $\mathbb{E}$ is the mathematical expectation function. $\mathcal{L}_{tr}$ is the training loss, $Perf_{\text{val}}$ is the validation performance, $\mathcal{D}_{tr}$ and $\mathcal{D}_{val}$ refer to the training and validation sets, respectively.

To optimize this process efficiently, we leverage LLM as an intelligent proxy $\mathcal{P}$ to guide the search process, enabling synergistic evolution of $\mathcal{A}_{aug}$ and a:

$$(a_{i+1}, \mathcal{A}_{aug,i+1}) = \mathcal{P}(\mathcal{S}_\alpha, \mathcal{D}_{\text{val}}, \delta(i), Perf_{\text{val}}(\delta(i)), \delta_0), \quad \text{s.t. } \beta(a) \leq \beta_0. \tag{6}$$

Here, a_{i+1} is the $(i+1)$-th iteration result, $\beta(a)$ denotes the architecture budget relative to β_0, and δ_0 represents all architecture and augmentation configurations. After each iteration, δ and $Perf_{\text{val}}(*)$ are updated:

$$Perf_{\text{val}}(\delta(i+1), \mathcal{D}_{\text{val}}) \leftarrow Perf(\delta(i), \mathcal{D}_{\text{val}}) + Perf(a_{i+1}, \mathcal{A}_{aug,i+1}, \mathcal{D}_{\text{val}}). \tag{7}$$



	Base Model	Ours	Ground Truth
(a)	On the superficial epithelium, tumor tissue consisting of proliferative images of medium-sized to large-sized round or irregular glandular ducts.	In the superficial epithelium, tumor tissue that invades by forming medium-sized to small, irregular ducts is observed moderately differentiated tubular adenocarcinoma.	In the superficial epithelium, tumor ti ssue that invades by forming medium -sized to small, irregular ducts is obse rved. moderately differentiated tubul ar adenocarcinoma.
(b)	Sections show a seborrhoeic keratosis there is acanthosis and hyperkeratosis with a proliferation of basaloid keratinocytes.	Sections show a seborrhoeic keratosis there is acanthosis and hyperkeratosis with a proliferation of basaloid keratinocytes without atypia and horn pseudocysts. no malignancy.	Sections show a seborrhoeic keratosis. There is acanthosis and hyperkeratosis with a proliferation of basaloid keratinocytes without atypia and horn pseudocysts. There is no evidence of malignancy.
(c)	The cardiomediastinal silhouette and pulmonary vasculature are within normal limits there is no pleural effusion or pneumothorax.	The cardiomediastinal silhouette and pulmonary vasculature are normal in size. The lungs are clear with no evidence of focal airspace disease, pneumothorax, or pleural effusion. No acute bony findings.	The cardiomediastinal silhouette and pulmonary vasculature are within normal limits in size. The lungs are clear of focal airspace disease, pneumothorax, or pleural effusion. There are no acute bony findings.

Fig. 4: Generated reports from (a) PatchGastric22, (b) Skin-Path, and (c) IU-Xray datasets, with abnormalities and correctly identified findings highlighted.

Consequently, the final result is selected from δ_i generated with Eq. 6:

$$\begin{aligned} (a^*, \mathcal{A}^*_{aug}) &= \underset{(a,\mathcal{A}_{aug})\in\delta}{\arg\max}\ Perf_{\text{val}}(a, \mathcal{A}_{aug} \mid \mathcal{D}_{val}), \\ \text{s.t. } w^*(a, \mathcal{A}_{aug}) &= \arg\min_{w} \mathcal{L}_{tr}(w, a, \mathcal{A}_{aug}; \mathcal{D}_{tr}). \end{aligned} \tag{8}$$

The proposed Dynamic Synergistic Evolution is outlined in Algorithm 1.

2.3 *Skin-Path*: The First VL Dataset for Skin Cancer

In the current medical field, Vision-Language (VL) datasets for skin cancer remain scarce. To fill the gap, we introduce *Skin-Path*, the first VL dataset for skin cancer, comprising 194 H&E-stained whole slide images (WSIs) from distinct patients at Southern Sun Pathology laboratory ($\times$20 magnification) with diagnostic reports by a senior dermatopathologist. From these WSIs, we extracted 277,761 patches of size 300$\times$300 pixel for MRG evaluation. Fig. 3(a) shows sample patches with their corresponding medical report. The dataset covers 10 common skin diseases, including seborrhoeic keratosis, basal cell carcinoma, and squamous cell carcinoma, enabling effective evaluation for automated skin cancer diagnosis. A word cloud in Fig. 3(b) illustrates the dataset's diversity.

3 Experiments

We evaluated UniMRG on three benchmarks: PatchGastric22 (262,777 patches from 991 WSIs) [24], Skin-Path, and IU-Xray (7,470 chest X-rays with 3,955 reports) [2]. Metrics included BLEU [19], METEOR (MTR) [3], ROUGE-L (RG-L) [12], and CIDEr (C) [25], using BLIP2 [9] as both VLFM and base model.

Implementation Details. Experiments used two NVIDIA RTX A6000 GPUs with ViT-L/14 (CLIP) [20] as image encoder and FlanT5 [9] as the language model. Training used a batch size of 16 and Adam optimizer (initial LR:

Table 2: Comparison results with pretrained models on the IU-Xray [2] dataset.

Model	Pretrain	BL-1	BL-2	BL-3	BL-4	MTR	RG-L
XrayGPT [23]	MIMIC+CheXpert	0.177	0.104	0.047	0.007	0.105	0.203
MiniGPT-4 [33]	MIMIC	0.389	0.262	0.181	0.134	0.169	0.308
Liu et al. [14]	MIMIC	0.499	0.323	0.238	0.184	0.208	0.390
Clinical-BERT [30]	MIMIC	0.495	0.330	0.231	0.170	-	0.376
UniMRG (Ours)	-	**0.503**	**0.336**	**0.252**	**0.196**	**0.228**	**0.415**

Table 3: Ablations on PatchGastric22, Skin-Path, and IU-Xray datasets.

Dataset	Model	BL-1	BL-2	BL-3	BL-4	MTR	RG-L	C	Δ
PatchGastric22	Base	0.460	0.382	0.325	0.285	0.256	0.449	1.850	-
	+UMA	0.478	0.403	0.349	0.311	0.265	0.465	2.284	8.1%
	+MCL	0.504	0.426	0.367	0.324	0.283	0.498	2.289	13.3%
	UniMRG (Ours)	**0.541**	**0.465**	**0.408**	**0.368**	**0.306**	**0.529**	**2.670**	**25.1%**
Skin-Path	Base	0.430	0.316	0.204	0.194	0.203	0.411	0.387	-
	+UMA	0.444	0.331	0.220	0.210	0.219	0.452	0.408	6.8%
	+MCL	0.456	0.351	0.245	0.233	0.232	0.476	0.419	13.7%
	UniMRG (Ours)	**0.478**	**0.384**	**0.268**	**0.256**	**0.246**	**0.533**	**0.438**	**22.9%**
IU-Xray	Base	0.462	0.299	0.202	0.150	0.172	0.341	0.329	-
	+UMA	0.471	0.305	0.217	0.162	0.184	0.362	0.344	5.3%
	+MCL	0.478	0.316	0.228	0.171	0.199	0.387	0.355	10.4%
	UniMRG (Ours)	**0.503**	**0.336**	**0.252**	**0.196**	**0.228**	**0.415**	**0.402**	**21.9%**

5e-5, exponential decay). Dataset and evaluation followed [24,30] for consistency. For DSE, the architecture search space includes FD/GC-Learner layers [1, 3], attention heads {2, 4, 8}, and hidden dimensions {128, 256, 512}.

3.1 Comparison with state-of-the-art methods

Table 1 compares UniMRG with state-of-the-art methods on the PatchGastric22 and IU-Xray datasets. UniMRG outperforms existing methods on nearly all metrics. Table 2 compares UniMRG with recent VLFM-based methods on IU-Xray, where it outperformed Clinical-BERT by 2.6% on BLEU-4 and 3.9% on ROUGE-L, demonstrating its ability to learn medically relevant features directly from the data without external knowledge or specialized training.

3.2 Ablation Studies

Effect of Dynamic Synergistic Evolution. Fig. 5 shows search results on PatchGastric22 and IU-Xray. Dashed lines mark baseline B4, MTR, and RL scores. Our method surpassed the baselines by iteration 3 on PatchGastric22 and iteration 6 on IU-Xray. Over 10 iterations, the best candidates were selected.

Impact of Different Components. We conducted ablations (Table 3), with full model achieving Avg.Δ gains of 25.1%, 22.9%, and 21.9%. Table 4 shows

Table 4: Impact of different data augmentation (DA) methods on benchmarks.

Dataset	**DA**	**BL-1**	**BL-2**	**BL-3**	**BL-4**	**MTR**	**RG-L**	**C**
	Mixup	0.507	0.429	0.376	0.328	0.286	0.501	2.224
	Cutout	0.512	0.436	0.380	0.330	0.288	0.492	2.309
PatchGastric22	Cutmix	0.516	0.438	0.394	0.342	0.290	0.508	2.394
	MCDA	0.524	0.453	0.395	0.354	0.293	0.515	2.554
	UMA	**0.541**	**0.465**	**0.408**	**0.368**	**0.306**	**0.529**	**2.670**
	Mixup	0.462	0.365	0.252	0.242	0.238	0.489	0.426
	Cutout	0.458	0.359	0.248	0.237	0.233	0.476	0.422
Skin-Path	Cutmix	0.465	0.368	0.255	0.248	0.240	0.509	0.428
	MCDA	0.469	0.372	0.259	0.250	0.242	0.516	0.431
	UMA	**0.478**	**0.384**	**0.268**	**0.256**	**0.246**	**0.533**	**0.438**
	Mixup	0.489	0.320	0.237	0.176	0.204	0.396	0.376
	Cutout	0.482	0.318	0.230	0.172	0.199	0.390	0.369
IU-Xray	Cutmix	0.491	0.325	0.244	0.181	0.210	0.402	0.388
	MCDA	0.495	0.328	0.247	0.186	0.215	0.407	0.393
	UMA	**0.503**	**0.336**	**0.252**	**0.196**	**0.228**	**0.415**	**0.402**

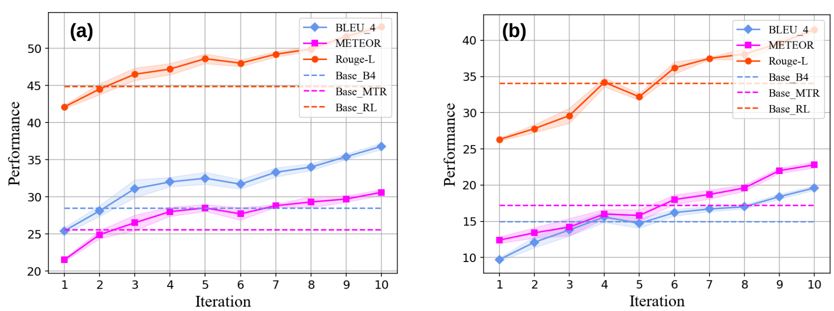


Fig. 5: Searching results for (a) PatchGastric22 and (b) IU-Xray datasets.

UMA outperforms Mixup [32], Cutout [4], Cutmix [31], and Manually Configured Data Augmentation (MCDA, fixed strategies without LLM guidance).

Impact of Module Integration on Model Size. We compare the number of parameters between the base model and our approach. Our method adds 2.95 million trainable parameters, a modest 2.86% increase over the base model.

Qualitative Analysis. Fig. 4 compares UniMRG-generated reports with the base model across three datasets. UniMRG more accurately captures disease details, producing reports that closely align with the ground truth.

4 Conclusion

In this work, we proposed UniMRG to address key MRG challenges by enhancing medical content perception and cross-modal integration via UMA and MCL. An LLM-guided evolution strategy jointly optimizes architecture and augmentation.

We also introduce the Skin-Path dataset covering 10 skin diseases. Experiments on PatchGastric22, IU-Xray, and Skin-Path confirm UniMRG's effectiveness.

5 Compliance with ethical standards

This study was performed in line with the principles of the Declaration of Helsinki. Ethics approval was granted CSIRO Health and Medical Human Research Ethics Committee (CHMHREC), under approval number 2021_030_LR, valid from 7 April 2021 to 7 April 2025. All experiments were conducted within the approval period, with no further data processing thereafter.

Acknowledgments. We thank Dr. Ian Katz for his valuable assistance with data annotation.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1439–1449 (2020)
2. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2016)
3. Denkowski, M., Lavie, A.: Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In: Proceedings of the sixth workshop on statistical machine translation. pp. 85–91 (2011)
4. DeVries, T.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
6. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
7. Jing, B., Wang, Z., Xing, E.: Show, describe and conclude: On exploiting the structure information of chest x-ray reports. arXiv preprint arXiv:2004.12274 (2020)
8. Jing, B., Xie, P., Xing, E.: On the automatic generation of medical imaging reports. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2577–2586 (2018)
9. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)

10. Li, Y., Liang, X., Hu, Z., Xing, E.P.: Hybrid retrieval-generation reinforced agent for medical image report generation. Advances in neural information processing systems **31** (2018)
11. Lin, C., Zhu, Z., Zhao, Y., Zhang, Y., He, K., Zhao, Y.: Sgt++: Improved scene graph-guided transformer for surgical report generation. IEEE Transactions on Medical Imaging **43**(4), 1337–1346 (2024). `https://doi.org/10.1109/TMI.2023.3335909`
12. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
13. Liu, A., Guo, Y., Yong, J.h., Xu, F.: Multi-grained radiology report generation with sentence-level image-language contrastive learning. IEEE Transactions on Medical Imaging (2024)
14. Liu, C., Tian, Y., Chen, W., Song, Y., Zhang, Y.: Bootstrapping large language models for radiology report generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 18635–18643 (2024)
15. Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y.: Exploring and distilling posterior and prior knowledge for radiology report generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13753–13762 (2021)
16. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
17. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
18. Pan, R., Ran, R., Hu, W., Zhang, W., Qin, Q., Cui, S.: S3-net: A self-supervised dual-stream network for radiology report generation. IEEE Journal of Biomedical and Health Informatics (2023)
19. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
20. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
22. Tan, M.: Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:1905.11946 (2019)
23. Thawkar, O., Shaker, A., Mullappilly, S.S., Cholakkal, H., Anwer, R.M., Khan, S., Laaksonen, J., Khan, F.S.: Xraygpt: Chest radiographs summarization using medical vision-language models. arXiv preprint arXiv:2306.07971 (2023)
24. Tsuneki, M., Kanavati, F.: Inference of captions from histopathological patches. In: International Conference on Medical Imaging with Deep Learning. pp. 1235–1250. PMLR (2022)
25. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)
26. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3156–3164 (2015)

27. Wang, L., Chen, J.: Improving radiology report generation with adaptive attention. In: Multimodal AI in healthcare: A paradigm shift in health intelligence, pp. 293–305. Springer (2022)
28. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pvt v2: Improved baselines with pyramid vision transformer. Computational Visual Media **8**(3), 415–424 (2022)
29. Xue, Y., Xu, T., Rodney Long, L., Xue, Z., Antani, S., Thoma, G.R., Huang, X.: Multimodal recurrent model with attention for automated radiology report generation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I. pp. 457–466. Springer (2018)
30. Yan, B., Pei, M.: Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2982–2990 (2022)
31. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6023–6032 (2019)
32. Zhang, H.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
33. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. In: ICLR (2024)