Multivariate Cooperative Game for Image-Report Pairs: Hierarchical Semantic Alignment for Medical Report Generation

Zhihong Zhu[1], Xuxin Cheng[1], Yunyan Zhang[2], Zhaorun Chen[3], Qingqing Long[4], Hongxiang Li[1], Zhiqi Huang[1], Xian Wu[2*], and Yefeng Zheng[2]

[1] School of Electronic and Computer Engineering, Peking University
[2] Jarvis Research Center, Tencent YouTu Lab
[3] Department of Computer Science, The University of Chicago
[4] Computer Network Information Center, Chinese Academy of Sciences
`zhihongzhu@stu.pku.edu.cn, kevinxwu@tencent.com`

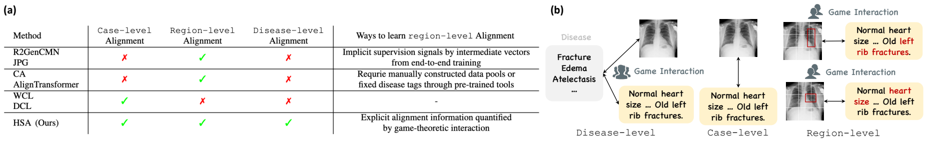
Abstract. Medical report generation (MRG) has great clinical potential, which could relieve radiologists from the heavy workloads of report writing. One of the core challenges in MRG is establishing accurate cross-modal semantic alignment between radiology images and their corresponding reports. Toward this goal, previous methods made great attempts to model from `case-level` alignment to more fine-grained `region-level` alignment. Although achieving promising results, they (1) either perform implicit alignment through end-to-end training or heavily rely on extra manual annotations and pre-training tools; (2) neglect to leverage the high-level inter-subject relationship semantic (*e.g.*, `disease`) alignment. In this paper, we present **H**ierarchical **S**emantic **A**lignment (HSA) for MRG in a unified game theory based framework, which achieves semantic alignment at multiple levels. To solve the first issue, we treat image regions and report words as binary game players and value possible alignment between them, thus achieving explicit and adaptive alignment in a self-supervised manner at `region-level`. To solve the second issue, we treat images, reports, and diseases as ternary game players, which enforces the cross-modal cluster assignment consistency at `disease-level`. Extensive experiments and analyses on IU-Xray and MIMIC-CXR benchmark datasets demonstrate the superiority of our proposed HSA against various state-of-the-art methods.

Keywords: Medical Report Generation · Multivariate Cooperative Game

1 Introduction

Medical report generation (MRG) aims to automatically generate coherent and informative reports to describe referring examination radiographs [20]. Due to its high potential in clinics, MRG has attracted extensive attention in recent years. Existing MRG works mainly adopt the encoder-decoder architecture, where the

[*] Corresponding Author


(a)
(b)
Disease
Fracture
Edema
Atelectasis
...
Game Interaction
Normal heart size ... Old left rib fractures.
Normal heart size ... Old left rib fractures.
Game Interaction
Normal heart size ... Old left rib fractures.
Game Interaction
Normal heart size ... Old left rib fractures.
Disease-level
Case-level
Region-level

Method	`Case-level` Alignment	`Region-level` Alignment	`Disease-level` Alignment	Ways to learn `region-level` Alignment
R2GenCMN JPG	✗	✓	✗	Implicit supervision signals by intermediate vectors from end-to-end training
CA AlignTransformer	✗	✓	✗	Requrie manually constructed data pools or fixed disease tags through pre-trained tools
WCL DCL	✓	✗	✗	-
HSA (Ours)	✓	✓	✓	Explicit alignment information quantified by game-theoretic interaction

Fig. 1: (a) Method comparison. (b) Our proposed HSA. (*Zoom-in for better view*)

encoder derives visual features from the input image and the decoder generates the corresponding report [1, 31, 34, 15]. The key challenge in MRG is how to accurately establish cross-modal semantic alignment between images and reports, which is crucial for precisely identifying complex diseases [40, 18, 42, 23].

To this end, a bunch of works attempted to establish from `case-level` alignment [38, 17] to `region-level` alignment [6, 41, 23, 40]. Therein, R2GenCMN [6] and JPG [41] treated memory vectors as the intermediary of vision and text modalities to enhance alignment in an end-to-end training manner. CA [23] compared the current input image with manually constructed normal image pool to obtain discriminative abnormal features. AlignTransformer [40] aligned image `region-level` features with several fixed disease tags from pre-training labelers [13]. Despite achieving promising progress, two main issues remain:

(1) Existing methods either achieve implicit alignment through end-to-end training or heavily rely on extra manual annotations and pre-training tools as summarized in Fig. 1(a). As a result, they often suffered from labor-intensive efforts or the risk of error propagation. This situation prompts a natural question: Can we achieve `region-level` alignment *explicitly and adaptively?* (2) They neglect to leverage the high-level inter-subject `disease-level` alignment. It is intuitive that unpaired images and reports sharing the same disease could also be semantically related [37], which is overlooked by most existing MRG works.

Following these premises, as shown in Fig.1(b), we propose **H**ierarchical **S**emantic **A**lignment (HSA) for MRG in a unified game theory based framework. Apart from vanilla `case-level` alignment, we introduce Banzhaf interaction [25] to achieve both `region-level` and `disease-level` alignment. The introduced game-theoretic interaction could mathematically value possible alignment among players at different levels, for explicit and adaptive alignment self-supervisedly.

To be specific, **to solve the first issue**, we formulate the image patches and report words in a pair as binary game players to perform for `region-level` alignment. Then, we utilize Banzhaf interaction [25] to value possible correspondence between image patches and report words for explicit and adaptive semantic alignment. **To solve the second issue**, we consider the images, reports and diseases as ternary game players to perform `disease-level` alignment. By valuing possible alignment between images/reports and diseases, we enforce the cross-modal cluster assignment consistency. We conduct extensive experiments to validate our framework on two benchmarks, IU-Xray [8] and MIMIC-CXR [14]. Empirical results and ablation studies show our method achieves significant improvements in almost all metrics that measure descriptive accuracy and clinical correctness.

Contributions. In a nutshell, the contributions of this work are three-fold:

- To our best knowledge, this is the first work to bring game theory into MRG. We present HSA, which utilizes game-theoretic interaction to achieve cross-modal semantic alignment at `disease-`, `case-` and `region`-level.
- To be specific, the proposed HSA performs different Banzhaf interactions at different levels correspondingly: image regions and report words (binary game) at `region-level`; images, reports, and diseases (ternary game) at `disease-level`. These two proposed interactions are unsupervised without labor-intensive effort, which could achieve explicit and adaptive alignment.
- Extensive experiments and analyses on IU-Xray and MIMIC-CXR benchmark datasets validate the effectiveness and superiority of the proposed HSA.

2 Preliminaries

2.1 Background of MRG

Given a medical image I, MRG aims to generate a radiology report $\hat{\mathbf{Y}}$ that describes findings $\mathbf{Y}$. Typical MRG systems are built upon encoder-decoder frameworks, which comprise an image encoder and a report decoder.

Image Encoder. In this work, we employ a Vision Transformer (ViT) [9] as our image encoder, which divides a medical image I into patches, with an additional `[CLS]` token to represent the global image feature. Through the image encoder, we obtain the encoded visual tokens $\mathbf{V} \in \mathbb{R}^{n_v \times d}$ and a global image representation $\mathbf{v}_g$, and $\mathbf{V}$ will be utilized for report generation.

Report Decoder. We adopt a two-layer standard Transformer decoder [32] as our report decoder. Technically, a [`Decode`] token is added to the beginning of $\mathbf{V}$ to signal the start while a [`EOS`] token is to signal its end. Through the report decoder, the output will be fed to a Linear & LogSoftmax layer to get the output of target sentences. Eventually, we train model parameters $\boldsymbol{\theta}$ to maximize $p_{\boldsymbol{\theta}}(\mathbf{Y}|I)$ by minimizing the negative log-likelihood loss: $\mathcal{L}_{\text{CE}} = -\sum_{t=1}^{\hat{T}} \log p_{\boldsymbol{\theta}}(\hat{y}_t|\hat{y}_{<t}, I)$, where $\hat{T}$ is the number of generated words.

2.2 Introduction of Game Theory

To achieve `region-level` and `disease-level` semantic alignment, we introduce the game theoretic interaction to mathematically value possible alignment at different levels. This subsection gives an introduction to game-theoretic interaction.

Notations. Specifically, the multivariate cooperative game consists of a set of players $\mathcal{P}$ with a revenue function $\phi(\mathcal{P})$. Therein, ϕ maps each team of players to a real score, which indicates the payoff obtained by $\mathcal{P}$ players working together to complete the task. The core of the game-theoretic interaction is to measure how much gain is obtained, and how to allocate the total gain fairly.

Interaction Strategy. In the multivariate cooperative game, there are various interaction strategies available, *i.e.*, core interaction [10], Shapley interaction [30] and Banzhaf interaction [25]. In this work, we choose Banzhaf interaction as our interaction strategy in MRG due to its balance of computational

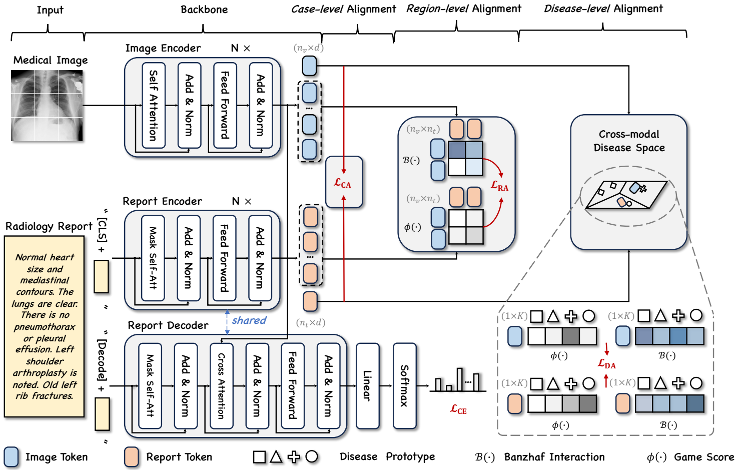


Fig. 2: The architecture of HSA. Beyond `case-level` alignment, we use Banzhaf interaction to value possible alignment between different players to achieve `region-level` and `disease-level` alignment. For simplicity, we only draw one image-report pair and several disease prototypes in `disease-level` alignment.

complexity and precision [12, 16]. Mathematically, given a coalition $\{i,j\} \subseteq \mathcal{P}$, the Banzhaf interaction $\mathcal{B}(\{i,j\})$ for the player $\{i,j\}$ is calculated as:

$$\mathcal{B}(\{i,j\}) = \sum_{\mathcal{C} \subseteq \mathcal{P} \setminus \{i,j\}} p(\mathcal{C})[\phi\left(\mathcal{C} \cup \{i,j\}\right) - \phi\left(\mathcal{C} \cup \{i\}\right) - \phi(\mathcal{C} \cup \{j\}) + \phi(\mathcal{C})], \tag{1}$$

where $p(\mathcal{C}) = \frac{1}{2^{n-2}}$ is the likelihood of the coalition $\mathcal{C}$ being sampled, and $\mathcal{P} \setminus \{i,j\}$ denotes removing $\{i,j\}$ from $\mathcal{P}$. $\mathcal{B}(\{i,j\})$ reflects the tendency of interactions inside $\{i,j\}$. The higher value of $\mathcal{B}(\{i,j\})$ indicates that player i and j cooperate closely with each other. For MRG, we take the matrix $\mathcal{B}$ as the **alignment label annotation**. We start with a detailed description of our HSA below.

3 Hierarchical Semantic Alignment

3.1 Vanilla Case-level Alignment

Following previous works, we first integrate the well-known `case-level` alignment. To be specific, we utilize a report encoder with a similar architecture to the image encoder to extract textual tokens $\mathbf{T} \in \mathbb{R}^{n_t \times d}$ and a global report representation $\mathbf{t}_g$. The loss function of `case-level` alignment can be formulated as two symmetric temperature-normalized InfoNCE [28] losses between global image and report representations in one mini-batch:

$$\mathcal{L}_{\mathrm{CA}} = -\frac{1}{2B}\left[\sum_{k=1}^{B} \log \frac{\exp\left(\hat{\mathbf{v}}_g^{(k)\top} \hat{\mathbf{t}}_g^{(k)} / \tau\right)}{\sum_{\ell}^{B} \exp\left(\hat{\mathbf{v}}_g^{(k)\top} \hat{\mathbf{t}}_g^{(\ell)} / \tau\right)} + \sum_{k=1}^{B} \log \frac{\exp\left(\hat{\mathbf{t}}_g^{(k)\top} \hat{\mathbf{v}}_g^{(k)} / \tau\right)}{\sum_{\ell}^{B} \exp\left(\hat{\mathbf{t}}_g^{(k)\top} \hat{\mathbf{v}}_g^{(\ell)} / \tau\right)}\right], \tag{2}$$

where B is the batch size; $\tau = 0.1$ is the temperature hyper-parameter; following common practice [5], $\hat{\mathbf{v}}_g^{(*)}$ and $\hat{\mathbf{t}}_g^{(*)}$ are normalized lower-dimensional global representations of $\mathbf{v}_g^*$ and $\mathbf{t}_g^*$, respectively.

3.2 Game Theory based Region-level Alignment

Since `case-level` alignment directly optimizes global representations for an image-report pair, they may miss important subtle clues. In contrast to prior works, we propose `region-level` alignment based on game theory, to achieve explicit and direct alignment. We first reiterate the players and define the game score ϕ_{RA}, then explain `region-level` semantic alignment in detail.

From the game-theoretic view, we take $\mathcal{M} = \{\mathbf{v}_i\}_{i=1}^{n_v} \cup \{\mathbf{t}_j\}_{j=1}^{n_t}$ as players. For the game score, we first define the alignment matrix: $\mathrm{A} = [a_{ij}]^{n_v \times n_t}$, where $a_{ij} = \mathbf{v}_i^\top \mathbf{t}_j$ represents the alignment score between i-th image patch and j-th report word. Next, $\tilde{\mathrm{A}}$ is obtained by applying row normalization of A. For the i-th image patch, we calculate its maximum alignment score as $\max_j \tilde{a}_{ij}$. Then, we use the weighted average maximum alignment score over all image patches as the image-to-report similarity $\mathbf{s}_1$. Similarly, we can obtain the report-to-image similarity $\mathbf{s}_2$. The total similarity score can be defined: $\mathbf{s} = (\mathbf{s}_1 + \mathbf{s}_2)/2$, which is considered as the game score $\phi_{\mathrm{RA}}(\mathcal{M})$ in `region-level` alignment.

Intuitively, if an image token has strong semantic correspondence with a report token, then they tend to cooperate with each other and contribute to the total game score. For a coalition $\{\mathbf{v}_i, \mathbf{t}_j\}$, referring to Eq. 1, we can calculate its Banzhaf interaction score $\mathcal{B}(\{\mathbf{v}_i, \mathbf{t}_j\})$. Then, we take normalized $\mathcal{B}'(\{\mathbf{v}_i, \mathbf{t}_j\})$ as **soft labels**, the loss function of `region-level` alignment can be formulated as:

$$\mathcal{L}_{\mathrm{RA}} = -\frac{1}{n_v n_t} \sum_{i=1}^{n_v} \sum_{j=1}^{n_t} \mathcal{B}'(\{\mathbf{v}_i, \mathbf{t}_j\}) \log(\tilde{a}_{ij}). \tag{3}$$

3.3 Game Theory based Disease-level Alignment

We devise `disease-level` alignment to harness the inter-subject alignment between images and reports, which imposes constraints on the embedding space, leading to the incorporation of more high-level semantic information.

Concretely, we first pre-define K trainable cross-modal disease prototypes $\mathcal{C} = \{\mathbf{c}_1, \cdots, \mathbf{c}_K\}$, where $\mathbf{c}_k \in \mathbb{R}^d$. After that, we calculate the visual softmax probability $\mathbf{p}_v \in \mathbb{R}^K$ of the cosine similarities between normalized global visual representation $\hat{\mathbf{v}}_g$ and cross-modal disease prototypes $\mathcal{C}$, and the text softmax probability $\mathbf{p}_t \in \mathbb{R}^K$ of the cosine similarities between normalized global textual representation $\hat{\mathbf{t}}_g$ and cross-modal disease prototypes $\mathcal{C}$. Then, we formulate $\hat{\mathbf{v}}_g$, $\hat{\mathbf{t}}_g$ and $\mathcal{C}$ as players in `disease-level` alignment. $\mathbf{p}_t$ is treated as the cross-modal game score between $\hat{\mathbf{v}}_g$ and $\mathcal{C}$ while $\mathbf{p}_v$ is treated as another cross-modal game score between $\hat{\mathbf{t}}_g$ and $\mathcal{C}$. The optimization is achieved by conducting cross-modal game-theoretic interaction of $\{\hat{\mathbf{v}}_g, \mathcal{C}\}$ and $\{\hat{\mathbf{t}}_g, \mathcal{C}\}$:

$$\mathcal{L}_{\mathrm{DA}} = \frac{1}{2} \left[\sum_{k=1}^{K} \mathcal{B}' \left(\{\hat{\mathbf{v}}_g, \mathbf{c}_k\}\right) \log \mathbf{p}_{k,t} + \sum_{k=1}^{K} \mathcal{B}' \left(\{\hat{\mathbf{t}}_g, \mathbf{c}_k\}\right) \log \mathbf{p}_{k,v} \right], \tag{4}$$

Table 1: Main results on IU-Xray and MIMIC-CXR datasets. ‡: our own re-implementation of baselines. *: our results significantly surpass baselines using paired t-test [39] with $p < 0.05$. –: missing results from the published work. Results in gray denote the model using different architectures.

Dataset	Method	NLG Metrcis							CE Metrcis		
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	Precision	Recall	F1
IU-Xray [8]	R2Gen [7]	0.470	0.304	0.219	0.165	0.187	0.371	-	-	-	-
	PPKED[22]	0.483	0.315	0.224	0.168	0.190	0.376	0.351	-	-	-
	R2GenCMN‡ [6]	0.471	0.306	0.222	0.170	0.193	0.374	0.343	-	-	-
	CA [23]	0.492	0.314	0.222	0.169	0.193	0.381	-	-	-	-
	CMCL [21]	0.473	0.305	0.217	0.162	0.186	0.378	-	-	-	-
	AlignTransformer [40]	0.484	0.313	0.225	0.173	0.204	0.379	-	-	-	-
	JPG [41]	0.479	0.319	0.222	0.174	0.193	0.377	-	-	-	-
	MSAT [36]	0.481	0.316	0.226	0.171	0.190	0.372	0.394	-	-	-
	MMTN [4]	0.486	0.321	0.232	0.175	-	0.375	0.361	-	-	-
	DCL [17]	-	-	-	0.163	0.193	0.383	0.586	-	-	-
	HSA (Ours)	**0.527***	**0.361***	**0.268***	**0.196***	**0.210***	**0.405***	**0.598***	-	-	-
	CvT-212DistilGPT2 [26]	0.473	0.304	0.224	0.175	0.200	0.376	0.694	-	-	-
	METransformer [35]	0.483	0.322	0.228	0.172	0.192	0.380	0.435	-	-	-
MIMIC-CXR [14]	R2Gen [7]	0.353	0.218	0.145	0.103	0.142	0.270	-	0.333	0.273	0.276
	PPKED [22]	0.360	0.224	0.149	0.106	0.149	0.284	0.237	-	-	-
	R2GenCMN‡ [6]	0.350	0.214	0.144	0.103	0.139	0.271	0.158	0.334	0.275	0.158
	M2TR [27]	0.378	0.232	0.154	0.107	0.145	0.272	-	0.240	**0.428**	0.308
	CA [23]	0.350	0.219	0.152	0.109	0.151	0.283	-	0.352	0.298	0.303
	CMCL [21]	0.344	0.217	0.140	0.097	0.133	0.281	-	-	-	-
	AlignTransformer [40]	0.378	0.235	0.156	0.112	0.158	0.283	-	-	-	-
	MSAT [36]	0.373	0.235	0.162	**0.120**	0.143	0.282	**0.299**	-	-	-
	WCL [38]	0.373	-	-	0.107	0.144	0.274	-	0.385	0.274	0.294
	MMTN [4]	0.379	0.238	0.159	0.116	0.161	0.283	-	-	-	-
	DCL [17]	-	-	-	0.109	0.150	0.284	0.281	0.471	0.352	0.373
	HSA (Ours)	**0.386***	**0.243***	**0.165***	**0.120**	**0.163***	**0.288***	0.287	**0.480***	0.357	**0.379**
	CvT-212DistilGPT2 [26]	0.393	0.248	0.171	0.127	0.155	0.286	0.389	0.367	0.418	0.391
	METransformer [35]	0.386	0.250	0.169	0.124	0.152	0.291	0.362	0.364	0.309	0.311

where the Banzhaf interaction $\mathcal{B}'(\{\hat{\mathbf{v}}_g, \mathbf{c}_k\})$ and $\mathcal{B}'(\{\hat{\mathbf{t}}_g, \mathbf{c}_k\})$ can be obtained using Eq.1 followed by normalization.

Finally, the overall training objective can be calculated as:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \left(\mathcal{L}_{\text{CA}} + \mathcal{L}_{\text{RA}} + \mathcal{L}_{\text{DA}}\right), \tag{5}$$

where λ is the trade-off hyper-parameter. In particular, HSA is only used during the training process to improve the representation learning, which can be directly removed during inference, rendering an efficient and semantics-sensitive model.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate our HSA on two widely-used MRG benchmarks, IU-Xray [8] and MIMIC-CXR [14]. We adopt the settings in [6, 22] to preprocess the reports for a fair comparison. **IU-Xray** comprises 3,955 radiology reports and 7,470 chest X-ray images. Following previous works [15, 6, 41], we apply the same split, *i.e.* 70%/10%/20%, for training/validation/test set. **MIMIC-CXR** is the largest released radiology dataset to date, which contains 377,110 radiographs

Table 2: Ablation study on the IU-Xray dataset. The **BASE** in (a) consists of an image encoder and a report decoder with $\mathcal{L}_{\mathrm{CE}}$ only. **CA**: `case-level` alignment. **RA**: `region-level` alignment. **DA**: `disease-level` alignment. **Training Time** in (b) denotes the average training time (s) for an epoch.

(a) Effect of each component in the proposed HSA.

Method	BL-4	MTR	RG-L	CDr
BASE	0.135	0.172	0.364	0.523
BASE+CA	0.174	0.193	0.389	0.567
BASE+CA+RA	0.187	0.204	0.398	0.586
HSA	**0.196**	**0.210**	**0.405**	**0.598**

(b) Effect of alignment strategy. CS: cosine similarity. GI: game-theoretic interaction.

Alignment	Strategy	BL-4	MTR	RG-L	CDr	Training Time↓
CA	-	0.174	0.193	0.389	0.567	
RA&DA	CS	0.189	0.204	0.401	0.589	**155**
	GI	**0.196**	**0.210**	**0.405**	**0.598**	161

(c) Effect of the initialization on image and report encoders.

ViT	SciBERT	BL-4	MTR	RG-L	CDr
✓	✗	0.182	0.205	0.393	0.582
✗	✓	0.180	0.196	0.381	0.554
✓	✓	**0.196**	**0.210**	**0.405**	**0.598**
✓	✓(BERT)	0.192	0.207	0.396	0.593

and 227,835 corresponding reports. We adopt the official splits [40], resulting in 368,960/2,991/5,159 in the training/validation/test set.

Metrics. We adopt natural language generation metrics (NLG Metrics) and clinical efficacy (CE Metrics) to evaluate HSA. BLEU [29], METEOR [2], ROUGE-L [19] and CIDEr [33] are selected as NLG Metrics, and we utilize the MS-COCO caption evaluation tool[¶] to calculate scores. For CE Metrics, we employ CheXpert[‖] proposed in [11] to label the generated reports and then compare with 14 disease labels of the references. Note that F1 in Table 1 refers to example-based macro F1 following [17].

The reported results are averaged over 5 runs with different seeds. Please refer to the supplementary material for more implementation details.

4.2 Comparison with State-of-the-arts

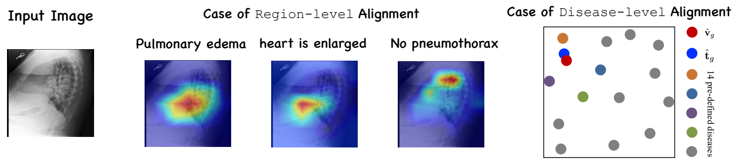
As shown in Tab.1, our HSA outperforms the baselines on almost all NLG metrics. This verifies the effectiveness of our HSA at three levels. However, on the MIMIC-CXR, our model is slightly inferior to MSAT (-1.2% CIDEr). This could be partly explained by MSAT employing a more powerful pre-trained image encoder (CLIP vs. ViT in ours) and a more potent attention mechanism (bilinear attention vs. vanilla attention in ours). Besides, our model achieves competitive results against previous SOTA methods in terms of CE Metrics. Specifically, our model achieves the best F1 score of 0.379, increasing by 0.6% compared to the best baseline. This indicates that our model can produce high-quality descriptions for clinical abnormalities.

4.3 Quantitative Analysis

We first explore whether each component contributes to the overall performance. As shown in Tab.2a, all the proposed `case-level` alignment (CA),

[¶] https://github.com/tylin/coco-caption

[‖] https://github.com/stanfordmlgroup/chexpert-labeler



Ground Truth: The heart is moderately enlarged. There is a widespread interstitial abnormality with indistinct pulmonary vascularity and upper zone redistribution, most consistent with moderate pulmonary edema. There is no definite pleural effusion or pneumothorax.

Baseline: Lung volumes are low resulting in bronchovascular crowding . There is no convincing evidence for pneumonia. However aspiration cannot be excluded . Prominent pulmonary vascularity is not significantly changed from. The heart is not enlarged.

Ours: Prominent pulmonary vascularity is indistinct. In the appropriate clinical setting, Pulmonary edema cannot be excluded. The heart is enlarged. No pneumothorax or pleural effusion.

Fig. 3: Case study of our proposed HSA and previous SOTA method DCL. At the top, we show the attention heatmaps of specific descriptions, and the disease embedding space by t-SNE [24]. We color the diseases and corresponding keywords in the report involved in this case, while keeping other diseases gray. Underlined text denotes the generated wrong sentences. (*Color figure online*)

`region-level` alignment (RA), and `disease-level` alignment (DA) benefit the performance of MRG. More quantitative analyses are presented below.

Effect of Alignment Strategy In Tab.2b, we explore the benefits of different alignment strategies. Comparing line #1 and line #2, the proposed RA and DA yield $>$ 0.1 % benefits from CA across four metrics. For the alignment strategy of RA and DA, our GI alignment significantly outperforms the intuitive CS alignment. Since the HSA can be removed during inference, we only consider training time for comparison as shown in Tab.2b. The results indicate minimal difference with 6s per epoch between the two strategies, whereas our game-theoretic interaction demonstrates a substantial performance improvement.

Effect of Encoder Initialization In our implementation, we utilize ViT [9] as image encoder and SciBERT [3] as report encoder considering the domain gap between medical and generic texts. Not surprisingly, the performances drop steeply without pre-trained ViT parameters when comparing line #2 and line #3 (*e.g.*, 0.598 → 0.554 on CIDEr). Moreover, comparison of the line #1, line #3 (w/ SciBERT) and line #4 (w/ BERT) demonstrates the impact of pre-trained SciBERT parameters and its tokenizer's sensitivity to medical terminology.

4.4 Qualitative Analysis

In Fig.3, we show a case from MIMIC-CXR to better understand our model. For visualizing the `region-level` alignment, we show the three attention heatmaps between the image clustering center and the report clustering center. It demonstrates that our proposed HSA can capture accurate fine-grained semantic alignment between image regions and keywords. For visualizing the `disease-level` alignment, we present the learned disease embedding space. It is observed that our HSA can learn reasonable `disease-level` semantic information. In contrast to previous SOTA method DCL, our HSA can generate more informative and meaningful radiology reports by enhancing hierarchical semantic alignment.

5 Conclusion

In this work, we made the first attempt to introduce multivariate cooperative game theory by formulating image-report as players at different levels. Based on this, we presented a hierarchical semantic alignment framework termed HSA, which seamlessly unifies the `case-level`, `region-level`, and `disease-level` alignment. Extensive experiments and analyses on two benchmark datasets demonstrated the superiority of our model over state-of-the-art methods.

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

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR (2018)
2. Banerjee, S., Lavie, A.: METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: ACL (2005)
3. Beltagy, I., Lo, K., Cohan, A.: Scibert: A pretrained language model for scientific text. In: EMNLP (2019)
4. Cao, Y., Cui, L., Zhang, L., Yu, F., Li, Z., Xu, Y.: Mmtn: Multi-modal memory transformer network for image-report consistent medical report generation. In: AAAI (2023)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.E.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
6. Chen, Z., Shen, Y., Song, Y., Wan, X.: Cross-modal memory networks for radiology report generation. In: ACL/IJCNLP (2021)
7. Chen, Z., Song, Y., Chang, T., Wan, X.: Generating radiology reports via memory-driven transformer. In: EMNLP (2020)
8. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S.K., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. J. Am. Medical Informatics Assoc. (2016)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
10. D'Orsogna, M.R., Chuang, Y.L., Bertozzi, A.L., Chayes, L.S.: Self-propelled particles with soft-core interactions: patterns, stability, and collapse. Physical review letters (2006)

11. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R.L., Shpanskaya, K.S., Seekins, J., Mong, D.A., Halabi, S.S., Sandberg, J.K., Jones, R., Larson, D.B., Langlotz, C.P., Patel, B.N., Lungren, M.P., Ng, A.Y.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: AAAI (2019)
12. Jin, P., Huang, J., Xiong, P., Tian, S., Liu, C., Ji, X., Yuan, L., Chen, J.: Video-text as game players: Hierarchical banzhaf interaction for cross-modal representation learning. In: CVPR (2023)
13. Jing, B., Xie, P., Xing, E.P.: On the automatic generation of medical imaging reports. In: ACL (2018)
14. Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C., Mark, R.G., Horng, S.: MIMIC-CXR: A large publicly available database of labeled chest radiographs. CoRR (2019)
15. Li, C.Y., Liang, X., Hu, Z., Xing, E.P.: Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In: AAAI (2019)
16. Li, H., Cao, M., Cheng, X., Li, Y., Zhu, Z., Zou, Y.: G2l: Semantically aligned and uniform video grounding via geodesic and game theory. In: CVPR (2023)
17. Li, M., Lin, B., Chen, Z., Lin, H., Liang, X., Chang, X.: Dynamic graph enhanced contrastive learning for chest x-ray report generation. In: CVPR (2023)
18. Li, Y., Yang, B., Cheng, X., Zhu, Z., Li, H., Zou, Y.: Unify, align and refine: Multi-level semantic alignment for radiology report generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
19. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out (2004)
20. Liu, C., Tian, Y., Song, Y.: A systematic review of deep learning-based research on radiology report generation. arXiv (2023)
21. Liu, F., Ge, S., Wu, X.: Competence-based multimodal curriculum learning for medical report generation. In: ACL/IJCNLP (2021)
22. Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y.: Exploring and distilling posterior and prior knowledge for radiology report generation. In: CVPR (2021)
23. Liu, F., Yin, C., Wu, X., Ge, S., Zhang, P., Sun, X.: Contrastive attention for automatic chest x-ray report generation. In: ACL/IJCNLP (2021)
24. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research (2008)
25. Marichal, J., Mathonet, P.: Weighted banzhaf power and interaction indexes through weighted approximations of games. Eur. J. Oper. Res. (2011)
26. Nicolson, A., Dowling, J., Koopman, B.: Improving chest x-ray report generation by leveraging warm starting. Artificial intelligence in medicine (2023)
27. Nooralahzadeh, F., Gonzalez, N.P., Frauenfelder, T., Fujimoto, K., Krauthammer, M.: Progressive transformer-based generation of radiology reports. In: EMNLP (2021)
28. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. CoRR (2018)
29. Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: a method for automatic evaluation of machine translation. In: ACL (2002)
30. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research (2020)
31. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: CVPR (2017)

32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
33. Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: Consensus-based image description evaluation. In: CVPR (2015)
34. Wang, X., Peng, Y., Lu, L., Lu, Z., Summers, R.M.: Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In: CVPR (2018)
35. Wang, Z., Liu, L., Wang, L., Zhou, L.: Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In: CVPR (2023)
36. Wang, Z., Tang, M., Wang, L., Li, X., Zhou, L.: A medical semantic-assisted transformer for radiographic report generation. In: MICCAI (2022)
37. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. In: EMNLP (2022)
38. Yan, A., He, Z., Lu, X., Du, J., Chang, E.Y., Gentili, A., McAuley, J.J., Hsu, C.: Weakly supervised contrastive learning for chest x-ray report generation. In: EMNLP (2021)
39. Yang, Y., Liu, X.: A re-examination of text categorization methods. In: SIGIR (1999)
40. You, D., Liu, F., Ge, S., Xie, X., Zhang, J., Wu, X.: Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. MICCAI (2022)
41. You, J., Li, D., Okumura, M., Suzuki, K.: JPG - jointly learn to align: Automated disease prediction and radiology report generation. In: COLING (2022)
42. Zhu, Z., Zhang, Y., Cheng, X., Huang, Z., Xu, D., Wu, X., Zheng, Y.: Alignment before awareness: Towards visual question localized-answering in robotic surgery via optimal transport and answer semantics. In: COLING (2024)