RRG-MLVA: RADIOLOGY REPORT GENERATION WITH MULTI-LEVEL VISION LANGUAGE ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Radiology report generators that couple a vision encoder with a large language model learn the correspondence between images and reports only implicitly, through the language-modelling loss. We propose RRG-MLVA, which connects a LoRA-adapted Swin Transformer to a LoRA-adapted LLaMA-2 decoder through a gated visual projection and trains them with four alignment objectives at complementary granularities: a global contrastive objective, a token-level payoff alignment in which report tokens respond to visual tokens through a quantal response, a multi-scale prototype objective, and a disease-level objective anchored to report-derived labels. The prototype objective treats learnable prototypes as players in a cooperative game whose coalition worth is the smooth maximum of their similarities to the input. Because this worth is non-additive, the Banzhaf value of a prototype measures what it adds beyond the other prototypes rather than its raw similarity; the values are estimated without bias by coalition sampling, and image and report must agree on them at three granularities. All objectives are used only during training, so the inference cost is unchanged. We evaluate on IU-Xray and a CheXpert Plus with language-generation metrics and CheXbert and RadGraph clinical-efficacy measures. RRG-MLVA reaches a CIDEr of 0.771 and a CheXbert F1 of 0.496 on IU-Xray, and a BLEU-4 of 0.350 and a CheXbert F1 of 0.440 on CheXpert Plus.

1 INTRODUCTION

Chest radiography is the most common imaging examination, and writing its free-text report is a large part of radiologists workload. Radiology report generation (RRG) aims to draft such reports automatically (Jing et al., 2018; Chen et al., 2020). Unlike generic captioning, a report must state the presence and absence of many findings, most of which are subtle deviations from normal anatomy, so a useful generator has to ground each statement in visual evidence rather than reproduce frequent normal phrases.

Recent generators couple a pre-trained vision encoder with a large language model (LLM) through a lightweight projection and adapt both with parameter-efficient fine-tuning (Wang et al., 2023; 2024; 2025b; Zhang et al., 2025). They inherit the fluency of the LLM but align the two modalities only implicitly: the visual tokens are shaped solely by what helps next-token prediction, and a loss dominated by recurring normal sentences can be low even when individual findings are weakly grounded. Explicit alignment at several granularities (image and report, regions and words, shared clinical findings) is well established for pre-training medical encoders (Zhang et al., 2022; Huang et al., 2021; Wang et al., 2022), but has received less attention as a training signal for LLM-based generators (Zhu et al., 2024).

Prototype based alignment compares the modalities at an intermediate semantic level: an image and its report should be explained in the same way by shared learnable prototypes (Caron et al., 2020; Wang et al., 2022). Existing formulations describe an input by a softmax over its prototype similarities, which scores each prototype independently, so overlapping prototypes all receive full credit and redundancy is rewarded. We instead treat the prototypes as players of a cooperative game and score them with the Banzhaf value (Banzhaf, 1965), a prototype's average marginal contribution to coalitions of the others. This requires a non-additive coalition worth: with an additive worth, the Banzhaf value collapses to the raw similarity.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We propose RRG-MLVA, a report generator trained with multi-level vision–language alignment. On top of a LoRA-adapted Swin Transformer (Liu et al., 2021) and LLaMA-2 (Touvron et al., 2023) connected by a gated projection, RRG-MLVA optimizes four training-only objectives on the visual tokens received by the LLM: a global contrastive objective; a token-level payoff alignment in which report tokens respond to visual tokens through a quantal response; a multi-scale Banzhaf prototype alignment; and a disease-level alignment to the 14 CheXpert observations (Irvin et al., 2019). For the prototype game we use the smooth maximum of the members' similarities as coalition worth, which has diminishing marginal contributions and admits an unbiased Monte Carlo estimator of the Banzhaf values. Our contributions are:

- A multi-level alignment framework for LLM-based report generation that combines global, token-level, prototype-level and disease-level objectives and adds no computation at inference.
- A cooperative-game formulation of prototype alignment with a non-additive worth, whose Banzhaf values discount redundant prototypes and are estimated without bias by coalition sampling at three granularities.
- Experiments on IU-Xray (Demner-Fushman et al., 2016) and a subset of CheXpert Plus (Chambon et al., 2024), including a controlled comparison with the same generator trained without alignment and ablations of each objective and of the number of prototype scales.

2 Related Work

Radiology report generation. R2Gen (Chen et al., 2020) introduced a memory driven Transformer, later extended to cross-modal memory (Chen et al., 2021). Following multimodal LLMs (Li et al., 2023; Liu et al., 2023), R2GenGPT (Wang et al., 2023) connected a Swin encoder to LLaMA through a lightweight alignment module, a design adopted by many successors: R2GenCSR (Wang et al., 2024) retrieves context samples, AM-MRG (Wang et al., 2025b) stores disease-related regions in associative memories, R2GenKG (Wang et al., 2025a) injects a multi-modal knowledge graph, and EMRRG (Zhang et al., 2025) fine-tunes an X-ray Mamba encoder; other work adds uncertainty-aware reasoning (Wang et al., 2025c) or an anatomical ontology (Li et al., 2025). These methods improve the visual input or the decoder, but image report correspondence is still learned implicitly; RRG-MLVA keeps the standard pipeline and adds explicit, training-only alignment.

Vision–language alignment in medical imaging. ConVIRT (Zhang et al., 2022) brought contrastive image–text learning (van den Oord et al., 2018; Radford et al., 2021) to radiology. GLoRIA (Huang et al., 2021) adds word-to-region attention, FILIP (Yao et al., 2022) a token-wise late-interaction similarity, and MGCA (Wang et al., 2022) combines instance-, token- and prototype-based disease-level alignment in the spirit of SwAV (Caron et al., 2020). Our global and token-level objectives follow these ideas, and our prototype objective is closest to MGCA, with two differences: we use alignment as auxiliary supervision for a report generator rather than for encoder pre-training, and we score prototypes by Banzhaf values, which account for redundancy, rather than by independent similarities.

Cooperative game theory in machine learning. Shapley (Shapley, 1953) and Banzhaf (Banzhaf, 1965) values are widely used for feature attribution (Lundberg & Lee, 2017) and data valuation (Ghorbani & Zou, 2019; Wang & Jia, 2023). Game-theoretic interactions have also served as training signals for fine-grained alignment: LOUPE (Li et al., 2022) uses Shapley interactions between patches and words, and HBI (Jin et al., 2023) Banzhaf interactions between video frames and words, both predicted by auxiliary networks; HSA (Zhu et al., 2024) applies Banzhaf interactions to report generation. In contrast, we compute the Banzhaf value of each individual player, a learnable prototype, for a closed-form non-additive worth over at most 28 players, which allows unbiased estimation by direct coalition sampling without a learned approximator, and we control granularity through the number of prototypes.

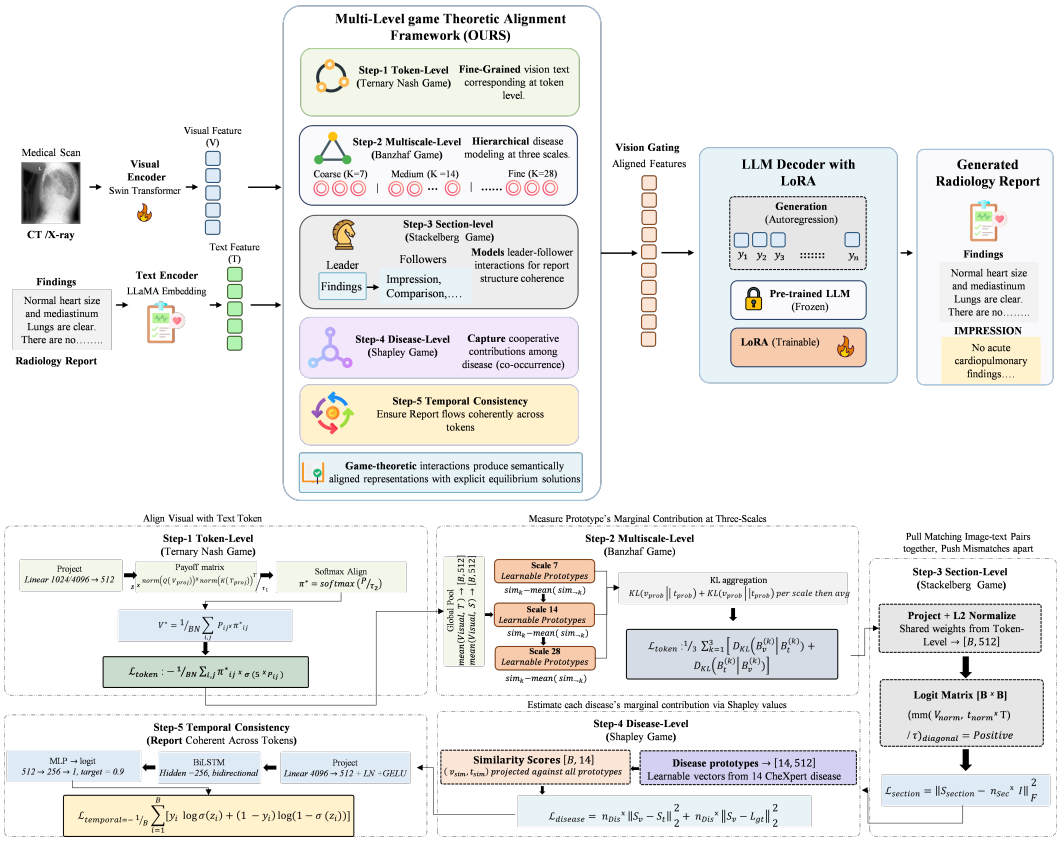

Figure 1: Overview of RRG-MLVA. Top: the report generator (LoRA-adapted Swin encoder, gated visual projection, LoRA-adapted LLaMA-2). Bottom: the training-only multi-level alignment module, which relates the projected visual tokens $\mathbf{H}$ to the embedded reference report $\mathbf{T}$ at the global, token, prototype and disease levels.

3 Method

3.1 Overview

Given a chest radiograph, or the radiographs $\mathcal{I} = \{I^{(1)}, \ldots, I^{(m)}\}$ of a sample, our goal is to generate the free-text report $\mathbf{y} = (y_1, \ldots, y_n)$. RRG-MLVA consists of two parts (Figure 1). The *report generator* (Section 3.2) encodes the images with a vision encoder, maps the visual tokens into the embedding space of a large language model (LLM) through a gated projection, and decodes the report autoregressively. The *multi-level alignment module* (Section 3.3) adds four auxiliary objectives that relate the visual tokens received by the LLM to the reference report at complementary granularities: a global contrastive objective, a token-level payoff alignment between report tokens and visual tokens, a multi-scale prototype alignment based on Banzhaf values of a cooperative game, and a disease-level alignment anchored to report-derived labels. The alignment module is used only during training and is discarded at inference, so decoding cost is identical to that of the underlying generator.

3.2 Report Generator

Visual encoding. Each image is encoded by a Swin Transformer (Liu et al., 2021) whose pre-trained weights are kept frozen and adapted with low-rank adapters (LoRA; Hu et al., 2022) inserted into the query and value projections of every self-attention layer. The encoder returns the patch tokens of its final stage, $\mathbf{X}^{(r)} \in \mathbb{R}^{N_v \times d_v}$ with $d_v = 1024$. When a sample contains several images, their token features are averaged, $\mathbf{X} = \frac{1}{m}\sum_{r=1}^{m} \mathbf{X}^{(r)}$.

Gated visual projection. The visual tokens are mapped to the LLM embedding dimension $d = 4096$ by a linear projection modulated by an input-dependent gate:

$$\mathbf{G} = \mathrm{Dropout}\big(\tanh(\mathrm{LN}(\mathbf{X}\mathbf{W}_g + \mathbf{b}_g))\big), \qquad \mathbf{H} = \mathrm{LN}\big((\mathbf{X}\mathbf{W}_p + \mathbf{b}_p) \odot (\mathbf{1} + \mathbf{G})\big), \tag{1}$$

where $\mathbf{W}_g, \mathbf{W}_p \in \mathbb{R}^{d_v \times d}$, LN denotes layer normalization, $\odot$ is the element-wise product and the dropout rate is 0.1. At inference, $\tanh(\cdot) \in (-1, 1)$ keeps the multiplicative factor $\mathbf{1} + \mathbf{G}$ in $(0, 2)$, and it reduces to the identity when the gate is inactive; the gate can therefore attenuate or amplify individual feature channels of each visual token without being able to suppress the visual signal entirely.

Prompted report generation. Following R2GenGPT (Wang et al., 2023), the visual tokens $\mathbf{H}$ are inserted into a fixed instruction prompt, and the LLM input sequence is

$$\big[\, \mathbf{e}_{\langle \mathrm{bos} \rangle};\ \mathbf{E}(\mathbf{p}_{\mathrm{pre}});\ \mathbf{H};\ \mathbf{E}(\mathbf{p}_{\mathrm{post}});\ \mathbf{E}(\mathbf{y}) \,\big], \tag{2}$$

where $\mathbf{E}$ is the LLM token-embedding table and $\mathbf{p}_{\mathrm{pre}}$, $\mathbf{p}_{\mathrm{post}}$ are the prompt segments before and after the visual tokens in the template

```
Human: <Img>[visual tokens]</Img> Generate a comprehensive and detailed
      diagnosis report for this chest xray image. Assistant:
```

We use LLaMA-2-7B-Chat (Touvron et al., 2023) as the decoder, with LoRA adapters in the query and value projections of every attention layer; all original LLM weights, including $\mathbf{E}$, remain frozen. The generator is trained with the standard next-token cross-entropy, computed only over report tokens (including the end-of-sequence token) and averaged over all such tokens in the mini-batch $\mathcal{B}$:

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{\sum_{b \in \mathcal{B}} n_b} \sum_{b \in \mathcal{B}} \sum_{t=1}^{n_b} \log p_\theta\big(y_t^{(b)} \mid y_{<t}^{(b)}, \mathbf{H}^{(b)}, \mathbf{p}\big). \tag{3}$$

3.3 Multi-Level Vision–Language Alignment

The cross-entropy objective aligns vision and language only implicitly, through what is useful for next-token prediction. We complement it with four explicit alignment objectives that operate on shared inputs.

Alignment inputs. On the visual side we use the projected visual tokens $\mathbf{H} = [\mathbf{h}_1; \ldots; \mathbf{h}_{N_v}] \in \mathbb{R}^{N_v \times d}$ of Eq. equation 1, i.e. exactly the tokens that the LLM receives. On the textual side, the reference report is tokenized to at most L tokens and embedded with the frozen LLM embedding table, $\mathbf{T} = \mathbf{E}(\tilde{\mathbf{y}}) = [\mathbf{t}_1; \ldots; \mathbf{t}_L] \in \mathbb{R}^{L \times d}$, with a binary mask $\mathbf{m} \in \{0, 1\}^L$ marking the $n = \sum_j m_j$ non-padding positions. Global descriptors are obtained by mean pooling over valid positions, $\bar{\mathbf{h}} = \frac{1}{N_v} \sum_i \mathbf{h}_i$ and $\bar{\mathbf{t}} = \frac{1}{n} \sum_j m_j \mathbf{t}_j$; padding positions are excluded from all alignment objectives. Because $\mathbf{T}$ is computed from frozen embeddings and the alignment objectives do not involve the LLM, their gradients reach the alignment heads, the gated visual projection and the LoRA adapters of the vision encoder, but not the LLM. The alignment objectives therefore shape the visual tokens that condition the decoder towards the content of the reference report, and they require no text at inference time. Throughout, $\cos(\mathbf{a}, \mathbf{b}) = \mathbf{a}^\top \mathbf{b} / (\|\mathbf{a}\| \, \|\mathbf{b}\|)$, and all alignment computations are carried out in single precision.

3.3.1 Global Contrastive Alignment

Two projection heads $\phi_v : \mathbb{R}^d \to \mathbb{R}^h$ and $\phi_t : \mathbb{R}^d \to \mathbb{R}^h$, each a two-layer MLP ($\mathrm{Linear} \to \mathrm{LN} \to \mathrm{GELU} \to \mathrm{Dropout} \to \mathrm{Linear} \to \mathrm{LN}$) with $h = 512$, map the global descriptors to a shared space, $\mathbf{u}_b = \phi_v(\bar{\mathbf{h}}_b)$ and $\mathbf{w}_b = \phi_t(\bar{\mathbf{t}}_b)$. With $s_{ab} = \cos(\mathbf{u}_a, \mathbf{w}_b)/\tau_c$ and a learnable temperature τ_c (lower-bounded by 10^{-2}), we use the symmetric InfoNCE objective (van den Oord et al., 2018; Radford et al., 2021; Zhang et al., 2022) over in-batch negatives:

$$\mathcal{L}_{\mathrm{con}} = -\frac{1}{2B} \sum_{b=1}^{B} \left[\log \frac{\exp(s_{bb})}{\sum_{b'} \exp(s_{bb'})} + \log \frac{\exp(s_{bb})}{\sum_{b'} \exp(s_{b'b})} \right], \tag{4}$$

where B is the number of image–report pairs in the micro-batch. We write $\mathrm{InfoNCE}(\mathbf{S})$ for this symmetric objective applied to a $B \times B$ logit matrix $\mathbf{S}$. The projection heads ϕ_v and ϕ_t are shared with the token-level objective below.

3.3.2 TOKEN-LEVEL PAYOFF ALIGNMENT

Findings in a report typically refer to localized image evidence, which a single global similarity cannot capture. We therefore relate every report token to every visual token. Using the heads ϕ_v, ϕ_t of Section 3.3.1, we form

$$\mathbf{q}_i = \mathbf{W}_q\, \phi_v(\mathbf{h}_i) + \mathbf{b}_q, \qquad \mathbf{k}_j = \mathbf{W}_k\, \phi_t(\mathbf{t}_j) + \mathbf{b}_k, \tag{5}$$

with $\mathbf{W}_q, \mathbf{W}_k \in \mathbb{R}^{h\times h}$. For a report a and an image b of the mini-batch, the payoff of report token j for visual token i is $P^{ab}_{ji} = \cos(\mathbf{k}^a_j, \mathbf{q}^b_i)$. Each report token answers with the logit (quantal) response (McKelvey & Palfrey, 1995) over visual tokens,

$$R^{ab}_{ji} = \frac{\exp(P^{ab}_{ji}/\tau_a)}{\sum_{i'=1}^{N_v} \exp(P^{ab}_{ji'}/\tau_a)}, \tag{6}$$

with a learnable temperature $\tau_a \in [10^{-2}, 1]$ that interpolates between a hard best response ($\tau_a \to 0$) and a uniform one. The expected payoff of token j under this response is $e^{ab}_j = \sum_i R^{ab}_{ji} P^{ab}_{ji}$, and averaging over the valid tokens of report a gives a pair score

$$S_{ab} = \frac{1}{n_a} \sum_{j=1}^{L} m^a_j\, e^{ab}_j. \tag{7}$$

Matching image–report pairs should obtain higher scores than non-matching ones, so we apply the symmetric InfoNCE objective to the logits $s\, S_{ab}$, where $s = \exp(\omega)$ is a learnable scale ($s \le 100$):

$$\mathcal{L}_{\text{tok}} = \text{InfoNCE}\big(s\, \mathbf{S}\big). \tag{8}$$

This late-interaction formulation is related to the word–region attention of GLoRIA (Huang et al., 2021) and the token-wise similarity of FILIP (Yao et al., 2022); using the batch as negatives prevents the trivial solution in which all tokens become equally similar.

3.3.3 MULTI-SCALE BANZHAF PROTOTYPE ALIGNMENT

Global and token-level objectives compare the two modalities directly. We additionally require that an image and its report be *explained* in the same way by a shared set of learnable semantic prototypes, and we measure how much each prototype explains an input with the Banzhaf value of a cooperative game (Banzhaf, 1965), following the use of Banzhaf interactions for cross-modal alignment (Jin et al., 2023) and for report generation (Zhu et al., 2024).

Prototypes and similarities. Linear maps $\psi_v : \mathbb{R}^d \to \mathbb{R}^h$ and $\psi_t : \mathbb{R}^d \to \mathbb{R}^h$ produce global features $\mathbf{f}^v = \psi_v(\bar{\mathbf{h}})$ and $\mathbf{f}^t = \psi_t(\bar{\mathbf{t}})$. At each scale $k \in \{1, 2, 3\}$ we maintain K_k learnable prototypes $\mathbf{C}^{(k)} = [\mathbf{c}_1; \ldots; \mathbf{c}_{K_k}] \in \mathbb{R}^{K_k \times h}$ with $(K_1, K_2, K_3) = (7, 14, 28)$, shared by both modalities. The prototypes are not tied to predefined categories; they are learned only through the objective below. For an input feature $\mathbf{f} \in \{\mathbf{f}^v, \mathbf{f}^t\}$ let $a_i = \cos(\mathbf{f}, \mathbf{c}_i)$.

Prototype game. At scale k the players are the prototypes $\mathcal{N} = \{1, \ldots, K_k\}$, and the worth of a coalition $S \subseteq \mathcal{N}$ is the smooth maximum of its members' similarities,

$$v(S) = \frac{1}{\beta} \log \sum_{i\in S} \exp(\beta a_i) \quad (S \neq \emptyset), \qquad v(\emptyset) = 0, \tag{9}$$

with sharpness $\beta > 0$ ($v(S) \to \max_{i\in S} a_i$ as $\beta \to \infty$). The Banzhaf value of prototype i is its marginal contribution averaged uniformly over all coalitions of the other players,

$$\varphi_i = \frac{1}{2^{K_k - 1}} \sum_{S \subseteq \mathcal{N}\setminus\{i\}} \big[v(S \cup \{i\}) - v(S)\big]. \tag{10}$$

The choice of a non-additive worth function is essential. If v were additive, $v(S) = \sum_{i\in S} a_i$, every marginal contribution would equal a_i and φ_i would reduce to the raw similarity. Under Eq. equation 9, the marginal contribution of i to a non-empty coalition is $\Delta_i(S) = \frac{1}{\beta} \log\big(1 + e^{\beta a_i} / \sum_{j\in S} e^{\beta a_j}\big)$, which is non-increasing in S: for non-empty $S \subseteq S' \subseteq \mathcal{N} \setminus \{i\}$ we have $\Delta_i(S) \ge \Delta_i(S')$. A prototype that duplicates what the coalition already covers therefore receives little credit, whereas a prototype that captures a pattern not covered by the others receives a large share, so redundant prototypes do not double-count evidence.

Monte Carlo estimation. Eq. equation 10 involves 2^{K_k-1} coalitions, which is intractable for $K_3 = 28$. Because the Banzhaf value weights all coalitions uniformly, it admits a simple unbiased estimator (Wang & Jia, 2023): we draw M membership masks $\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(M)} \in \{0,1\}^{K_k}$ with i.i.d. Bernoulli$(1/2)$ entries, set $S_i^{(m)} = \{j \neq i : z_j^{(m)} = 1\}$, and compute

$$\hat{\varphi}_i = \frac{1}{M} \sum_{m=1}^{M} \big[v(S_i^{(m)} \cup \{i\}) - v(S_i^{(m)}) \big]. \tag{11}$$

Each $S_i^{(m)}$ is uniformly distributed over the subsets of $\mathcal{N} \setminus \{i\}$, hence $\mathbb{E}[\hat{\varphi}_i] = \varphi_i$. The same M masks are reused for all players, and are shared between the visual and the textual game (common random numbers), so that differences between the two modalities are not dominated by sampling noise; the masks are redrawn at every training step, and the cost is $\mathcal{O}(MK_k^2)$ per sample and scale, negligible relative to the LLM. Because v depends only on the K_k precomputed similarities, evaluating a coalition requires no additional pass through the encoders; direct sampling is therefore affordable, and no learned approximation of the game is needed, unlike in token-level games whose interactions are predicted by auxiliary networks (Li et al., 2022; Jin et al., 2023).

Alignment objective. Let $\hat{\boldsymbol{\varphi}}_{v,b}^{(k)}, \hat{\boldsymbol{\varphi}}_{t,b}^{(k)} \in \mathbb{R}^{K_k}$ be the Banzhaf vectors of image b and report b at scale k. We combine three terms. (a) *Cross-sample matching:* with centred, ℓ_2-normalized vectors $\mathbf{z}_{v,b} = \mathrm{norm}(\hat{\boldsymbol{\varphi}}_{v,b} - \mathrm{mean}(\hat{\boldsymbol{\varphi}}_{v,b}))$ and $\mathbf{z}_{t,b}$ defined analogously, an image must be explained by the prototypes in the same way as its own report and differently from the other reports of the batch, $\mathcal{L}_{\mathrm{match}}^{(k)} = \mathrm{InfoNCE}\big(s_{\mathrm{ms}} \mathbf{Z}_v \mathbf{Z}_t^\top\big)$ with a fixed scale $s_{\mathrm{ms}} = 10$. (b) *Distribution agreement:* with $\boldsymbol{\pi} = \mathrm{softmax}(\hat{\boldsymbol{\varphi}}/\tau_{\mathrm{ms}})$ and a fixed temperature $\tau_{\mathrm{ms}} = 0.1$, $\mathcal{L}_{\mathrm{KL}}^{(k)} = \frac{1}{2B} \sum_b \big[D_{\mathrm{KL}}(\boldsymbol{\pi}_{t,b} \| \boldsymbol{\pi}_{v,b}) + D_{\mathrm{KL}}(\boldsymbol{\pi}_{v,b} \| \boldsymbol{\pi}_{t,b}) \big]$. (c) *Prototype diversity:* $\mathcal{L}_{\mathrm{div}}^{(k)} = \frac{1}{K_k(K_k-1)} \sum_{i \neq j} \cos(\mathbf{c}_i, \mathbf{c}_j)^2$ discourages prototypes from collapsing onto each other. The multi-scale objective averages over scales,

$$\mathcal{L}_{\mathrm{ms}} = \frac{1}{3} \sum_{k=1}^{3} \Big(\mathcal{L}_{\mathrm{match}}^{(k)} + \mathcal{L}_{\mathrm{KL}}^{(k)} + 0.1\, \mathcal{L}_{\mathrm{div}}^{(k)} \Big). \tag{12}$$

The temperature is fixed because a learnable one can make both distributions uniform, and the matching term supplies negatives, so the agreement term cannot be satisfied trivially. Using prototype sets of increasing size encourages agreement between the modalities on both coarse and fine partitions of the feature space; with fewer scales, the average in Eq. equation 12 runs over the scales that are used.

3.3.4 Disease-Level Alignment

The objectives above are self-supervised. To inject explicit clinical semantics we additionally align both modalities with the 14 CheXpert observation categories (Irvin et al., 2019). For each training report we derive a binary label vector $\mathbf{y}^{\mathrm{dis}} \in \{0,1\}^{14}$ with a lightweight rule-based extractor that matches category-specific keywords and discards mentions preceded by a negation cue (e.g., "no", "without", "negative for"). Linear maps $\rho_v : \mathbb{R}^d \to \mathbb{R}^h$ and $\rho_t : \mathbb{R}^d \to \mathbb{R}^h$ project $\bar{\mathbf{h}}$ and $\bar{\mathbf{t}}$, and a set of learnable disease prototypes $\mathbf{D} = [\mathbf{d}_1; \dots; \mathbf{d}_{14}] \in \mathbb{R}^{14 \times h}$ yields per-category scores $r_c^v = \cos(\rho_v(\bar{\mathbf{h}}), \mathbf{d}_c)$ and $r_c^t = \cos(\rho_t(\bar{\mathbf{t}}), \mathbf{d}_c)$. The objective anchors the textual scores to the labels and distills them into the visual scores:[1]

$$\mathcal{L}_{\mathrm{dis}} = \frac{1}{2} \cdot \frac{1}{14B} \big\| \mathbf{R}^v - \mathbf{R}^t \big\|_F^2 + \frac{1}{2} \cdot \frac{1}{14B} \big\| \mathbf{R}^t - \mathbf{Y}^{\mathrm{dis}} \big\|_F^2, \tag{13}$$

where $\mathbf{R}^v, \mathbf{R}^t \in [-1,1]^{B \times 14}$ stack the scores and $\mathbf{Y}^{\mathrm{dis}} \in \{0,1\}^{B \times 14}$ the labels of the batch. The visual branch is never supervised by the labels directly; it receives clinical supervision only through agreement with the label-anchored textual branch.

[1] The implementation also contains a regularizer on a co-occurrence matrix that is not used by any other part of the model; it has no effect on training and is omitted here.

3.4 Training Objective and Inference

The full model is trained end-to-end with

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{con}}\mathcal{L}_{\mathrm{con}} + \lambda_{\mathrm{tok}}\mathcal{L}_{\mathrm{tok}} + \lambda_{\mathrm{ms}}\mathcal{L}_{\mathrm{ms}} + \lambda_{\mathrm{dis}}\mathcal{L}_{\mathrm{dis}}, \tag{14}$$

with $(\lambda_{\mathrm{con}}, \lambda_{\mathrm{tok}}, \lambda_{\mathrm{ms}}, \lambda_{\mathrm{dis}}) = (0.6, 0.5, 0.4, 0.4)$ for all experiments. At inference the alignment module is removed and the report is decoded from the prompt and $\mathbf{H}$ alone with deterministic beam search (beam size 3, no sampling), 4-gram repetition blocking and length penalty 1.0. Because typical report lengths differ strongly between the datasets (about 90 words for CheXpert Plus and 30 words for IU-Xray in the training sets), the length limits are set per dataset: 80–150 new tokens with repetition penalty 2.5 for CheXpert Plus, and 10–80 new tokens with repetition penalty 1.5 for IU-Xray. The same decoding settings are used for validation and test and for all compared variants.

3.5 Implementation Details

The vision encoder is Swin-B (Liu et al., 2021) and the decoder LLaMA-2-7B-Chat (Touvron et al., 2023); both use LoRA (rank 8, $\alpha = 32$, dropout 0.1), and the gated projection and all alignment heads are trained from scratch. We use $h = 512$, $L = 512$, $M = 32$ coalitions and $\beta = 8$; τ_c and τ_a are initialized to 0.07 and 0.1, and the Banzhaf and disease prototypes from $\mathcal{N}(0, 0.02^2)$ and $\mathcal{N}(0, 0.01^2)$. We train for 30 epochs with AdamW (Loshchilov & Hutter, 2019) (learning rate 2×10^{-5}, weight decay 0.01), a per-epoch cosine schedule decaying to 10^{-6}, gradient clipping at 1.0 and bfloat16 precision, using micro-batches of $B = 4$, which also provide the in-batch negatives, with 8 gradient-accumulation steps. The checkpoint with the best validation score $0.5 \cdot$ BLEU-4 $+$ $0.5 \cdot$ CIDEr is evaluated once on the test set; all variants share the seed (42) and all settings except the ablated component.

4 Experiments

4.1 Experimental Setup

Datasets We evaluate RRG_MLGT on two standard chest X-ray report generation benchmarks: CheXpert Plus and IU-XRay. CheXpert Plus is a large-scale dataset containing 223,462 chest X-ray images paired with reports from 187,711 studies across 64,725 patients. Each report is organized into 11 clinical sections and annotated with 14 CheXpert-defined pathologies. We follow a patient-level split of 7:1:2 for training, validation, and testing to avoid data leakage.

We evaluate on IU-XRay, which includes 7,470 chest X-ray images and 3,955 associated radiology reports collected from the Indiana University hospital network. Each report contains four structured sections: Comparison, Indication, Findings, and Impression. For fair comparison, we use the standard train/validation/test splits provided in prior work.

Metrics. We report BLEU-1–4 (Papineni et al., 2002), ROUGE-L (Lin, 2004) and CIDEr (Vedantam et al., 2015) with the COCO caption toolkit and, for clinical correctness, the CheXbert, F1 over the 14 CheXpert observations (CheXb-14) and RadGraph F1 (Jain et al., 2021). All results come from a single training run per variant.

Compared methods. On IU-Xray we compare with published results of R2Gen (Chen et al., 2020), R2GenGPT (Wang et al., 2023), HSA (Zhu et al., 2024), R2GenKG (Wang et al., 2025a) and AM-MRG (Wang et al., 2025b). On both datasets we also train **Base**, the same generator with all alignment weights set to zero, which is the most direct reference for the effect of alignment.

4.2 Main Results

Table 1 places RRG-MLVA among published methods on IU-Xray, where it reaches a BLEU-4 of 0.340, a ROUGE-L of 0.482 and a CIDEr of 0.771, above all listed methods on every language metric. Because our image-level protocol differs from the study-level protocol of prior work, this comparison is indicative only; we therefore draw our conclusions from the evaluation in Table 2 and the ablation study in Table 3, in which all variants share the same splits, decoding settings, seed and checkpoint-selection rule.

Table 1: Comparison with published methods on IU-Xray. †Our model uses an image-level split that differs from the study-level split of prior work; the comparison is indicative only. Best value per column in bold.

Method	Venue	B-1	B-2	B-3	B-4	R-L	CIDEr
R2Gen (Chen et al., 2020)	EMNLP 2020	0.470	0.304	0.219	0.165	0.371	–
R2GenGPT (Wang et al., 2023)	Meta-Radiology 2023	0.488	0.316	0.228	0.173	0.377	0.438
HSA (Zhu et al., 2024)	MICCAI 2024	0.527	0.361	0.268	0.196	0.405	0.598
R2GenKG (Wang et al., 2025a)	arXiv 2025	0.468	0.312	0.231	0.181	0.383	0.701
AM-MRG (Wang et al., 2025b)	IEEE TMI 2025	0.489	0.339	0.253	0.192	0.384	0.613
RRG-MLVA (ours)†	–	**0.563**	**0.403**	**0.366**	**0.340**	**0.482**	**0.771**

Table 2: Performance evaluation of RRG-MLVA across IU-Xray and CheXpert Plus (subset). Checkpoints are selected on the validation set and evaluated once on the test set.

Dataset	B-1	B-2	B-3	B-4	R-L	CIDEr	CheXb-14	RadGraph
IU-Xray	0.5630	0.4030	0.3660	0.3400	0.4820	0.7712	0.4963	0.4475
CheXpert Plus (subset)	0.4939	0.4169	0.3770	0.3499	0.4773	0.1093	0.4404	0.1759

Table 2 presents the main evaluation results under identical training, decoding, and selection settings. RRG-MLVA reaches a BLEU-4 of 0.340, a CIDEr of 0.771, a CheXbert F1 of 0.496 and a RadGraph F1 of 0.448 on IU-Xray, and a BLEU-4 of 0.350, a CIDEr of 0.109, a CheXbert F1 of 0.440 and a RadGraph F1 of 0.176 on the CheXpert Plus subset. The much lower CIDEr on CheXpert Plus is consistent with its far longer reference reports (about 90 vs. about 30 words on IU-Xray), which the metric penalizes; the clinical metrics are therefore the primary evaluation focus on this dataset.

Clinical efficacy. On the CheXpert Plus subset, RRG-MLVA reaches a CheXbert F1 of 0.440 and a RadGraph F1 of 0.176; on IU-Xray, 0.496 and 0.448 (Table 2). Because our splits differ from those of prior work and clinical-efficacy scores depend on the labeler and the averaging protocol, we do not compare these values with published CE numbers; the detailed component interactions are evaluated via the ablation variants (Table 3).

4.3 Ablation Study

Table 3 compares the full model with variants that remove one objective by setting its weight to zero, and with variants that use fewer prototype scales, on both datasets. All rows are test-set results of checkpoints selected on the validation set; all variants share the seed and all settings except the ablated component.

Contribution of each objective. Removing any single objective degrades both language-generation and clinical metrics on both datasets, so no component is redundant. The disease-level objective, the only one supervised by clinical labels, contributes most to clinical efficacy: removing it lowers CheXbert F1 by 6.8 points on IU-Xray (0.4963→0.4281) and 6.0 points on CheXpert Plus (0.4404→0.3809), and RadGraph F1 by 3.6 and 1.1 points, respectively. The token-level objective contributes most to language quality: removing it causes the largest CIDEr drop on IU-Xray (0.7712→0.6065) and the largest RadGraph drop on CheXpert Plus (0.1759→0.1605). Removing the multi-scale prototype objective costs 2.4–3.1 CheXbert points and 1.5–2.3 RadGraph points across the two datasets, consistent with its role of discounting redundant prototypes so that image and report must agree on which patterns are covered. The global contrastive objective has the smallest effect on CheXbert F1 (-0.5 points on IU-Xray, -1.2 on CheXpert Plus) but the second-largest CIDEr drop on IU-Xray (0.7712→0.6074), suggesting that it mainly shapes the global organization of the report rather than individual findings.

Number of prototype scales. On CheXpert Plus, language scores are best with a single scale (BLEU-4 0.3562, ROUGE-L 0.4877, CIDEr 0.1190) and decrease slightly as scales are added, whereas clinical efficacy increases monotonically with the number of scales (CheXbert F1 0.3976

Table 3: Ablation study (test set). Each "w/o" row sets one weight to zero; the bottom rows use fewer prototype scales in $\mathcal{L}_{\mathrm{ms}}$ (the full model uses $\{7, 14, 28\}$). Best value per column in bold.

	CheXpert Plus (subset)					IU-Xray				
Variant	B-4	R-L	CIDEr	CheXb-14	RadGraph	B-4	R-L	CIDEr	CheXb-14	RadGraph
Full RRG-MLVA	0.3499	0.4773	0.1093	**0.4404**	**0.1759**	**0.3400**	**0.4820**	**0.7712**	**0.4963**	**0.4475**
w/o $\mathcal{L}_{\mathrm{con}}$ (global)	0.3388	0.4686	0.1021	0.4288	0.1670	0.3289	0.4343	0.6074	0.4909	0.4254
w/o $\mathcal{L}_{\mathrm{tok}}$ (token)	0.3391	0.4674	0.1026	0.4199	0.1605	0.3272	0.4288	0.6065	0.4648	0.4353
w/o $\mathcal{L}_{\mathrm{ms}}$ (prototype)	0.3347	0.4660	0.1005	0.4164	0.1608	0.3219	0.4235	0.6902	0.4655	0.4245
w/o $\mathcal{L}_{\mathrm{dis}}$ (disease)	0.3360	0.4643	0.1016	0.3809	0.1647	0.3347	0.4355	0.6759	0.4281	0.4113
One scale $\{14\}$	**0.3562**	**0.4877**	**0.1190**	0.3976	0.1758	–	–	–	–	–
Two scales $\{7, 14\}$	0.3506	0.4771	0.1000	0.4153	0.1728	–	–	–	–	–

for $\{14\}$, 0.4153 for $\{7, 14\}$ and 0.4404 for the full $\{7, 14, 28\}$), with RadGraph F1 essentially tied between one and three scales (0.1758 vs. 0.1759). The full three-scale setting thus trades a small amount of N-gram overlap for the best clinical grounding, which is the target of the alignment objectives.

5 Conclusion

We presented RRG-MLVA, an LLM-based report generator trained with training-only alignment objectives at the global, token, prototype and disease levels, whose prototype objective scores prototypes by Banzhaf values of a cooperative game with a non-additive worth. Our experiments use a CheXpert Plus subset and an image-level IU-Xray split, so absolute numbers are not directly comparable with prior work; each variant was trained once, so small differences should be read with care; the in-batch negatives come from micro-batches of four samples; and the training disease labels come from a rule-based extractor and are noisy. Generated reports can contain errors and are not suitable for clinical use without expert review.

AI USE STATEMENT

In this article, ChatGPT model 5.4 (OpenAI, CA, USA) was used for refining texts. All content was reviewed for accuracy and coherence.

ETHICS STATEMENT

This work uses the publicly available, de-identified IU-Xray and CheXpert Plus datasets under their respective data use agreements; no new patient data were collected. Automatically generated radiology reports may omit findings or describe findings that are not present, and a model trained on data from a small number of institutions may not generalize to other populations or imaging equipment. RRG-MLVA is a research prototype intended to assist, not replace, radiologists, and its outputs must be reviewed by qualified clinicians before any clinical use.

REPRODUCIBILITY STATEMENT

Section 3 specifies every objective, and Section 3.5 lists all architectural choices, hyper-parameters, initializations, the optimizer and the checkpoint-selection rule. Dataset splits, preprocessing and dataset-specific decoding settings are described in Sections 4.1 and 3.4. All variants in the ablation study differ only in the stated loss weights or prototype scales and share the same seed. We will release the code, the lists of sample identifiers of all splits, and the evaluation scripts upon publication.

REFERENCES

John F. Banzhaf, III. Weighted voting doesn't work: A mathematical analysis. *Rutgers Law Review*, 19:317–343, 1965.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9912–9924, 2020.

P. Chambon, J.-B. Delbrouck, T. Sounack, S.-C. Huang, Z. Chen, M. Varma, S. Q. Truong, C. T. Chuong, and C. P. Langlotz. CheXpert Plus: Augmenting a large chest X-ray dataset with text radiology reports, patient demographics and additional image formats. *arXiv preprint arXiv:2405.19538*, 2024.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 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.

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

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

Amirata Ghorbani and James Zou. Data Shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning (ICML)*, pp. 2242–2251, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3922–3931, 2021.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 590–597, 2019.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, Curtis P Langlotz, and Pranav Rajpurkar. RadGraph: Extracting clinical entities and relations from radiology reports. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

P. Jin, J. Huang, P. Xiong, S. Tian, C. Liu, X. Ji, L. Yuan, and J. Chen. Video-text as game players: Hierarchical Banzhaf interaction for cross-modal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2472–2482, 2023.

Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2577–2586, 2018.

Juncheng Li, Xin He, Longhui Wei, Long Qian, Linchao Zhu, Lingxi Xie, Yueting Zhuang, Qi Tian, and Siliang Tang. Fine-grained semantically aligned vision-language pre-training. In *Advances in Neural Information Processing Systems*, volume 35, pp. 7290–7303, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, pp. 19730–19742, 2023.

Q. Li, Z. Cui, S. Bae, J. Xu, R. Yuan, Y. Zhang, R. Feng, Q. Shen, X. Zhang, S. Gao, J. He, and S. Wang. AOR: Anatomical ontology-guided reasoning for medical large multimodal model in chest X-ray interpretation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 19331–19371, 2025.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10002, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Richard D. McKelvey and Thomas R. Palfrey. Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6–38, 1995.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 311–318, 2002.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Lloyd S. Shapley. A value for n-person games. In *Contributions to the Theory of Games II*, volume 28 of *Annals of Mathematics Studies*, pp. 307–317. Princeton University Press, 1953.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4566–4575, 2015.

F. Wang, Y. Qiao, X. Wang, F. Wang, Y. Zhang, and D. Sun. R2GenKG: Hierarchical multi-modal knowledge graph for LLM-based radiology report generation. *arXiv preprint arXiv:2508.03426*, 2025a.

Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 33536–33549, 2022.

Jiachen T. Wang and Ruoxi Jia. Data Banzhaf: A robust data valuation framework for machine learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

X. Wang, Y. Li, F. Wang, S. Wang, C. Li, and B. Jiang. R2GenCSR: Retrieving context samples for large language model based X-ray medical report generation. *arXiv preprint arXiv:2408.09743*, 2024.

X. Wang, F. Wang, H. Wang, B. Jiang, C. Li, Y. Wang, Y. Tian, and J. Tang. Activating associative disease-aware vision token memory for LLM-based X-ray report generation. *IEEE Transactions on Medical Imaging*, 2025b.

Z. Wang, S. Yan, K. Yin, X. Zhang, and W. K. Cheung. CURV: Coherent uncertainty-aware reasoning in vision-language models for X-ray report generation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 93686–93713, 2025c.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. R2GenGPT: Radiology report generation with frozen LLMs. *Meta-Radiology*, 1(3):100033, 2023.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: Fine-grained interactive language-image pre-training. In *International Conference on Learning Representations*, 2022.

M. Zhang et al. EMRRG: Efficient fine-tuning pre-trained X-ray Mamba networks for radiology report generation. *arXiv preprint arXiv:2510.16776*, 2025.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Proceedings of the Machine Learning for Healthcare Conference (MLHC)*, volume 182 of *Proceedings of Machine Learning Research*, pp. 2–25, 2022.

Z. Zhu, X. Cheng, Y. Zhang, Z. Chen, Q. Long, H. Li, Z. Huang, X. Wu, and Y. Zheng. Multivariate cooperative game for image-report pairs: Hierarchical semantic alignment for medical report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 303–313, Cham, 2024. Springer Nature Switzerland.