HOLISTIC DIAGNOSTIC SEMANTICS FROM DECOMPOSED IMAGING FOR ENDOSCOPIC REPORT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Endoscopic Report Generation (ERG) aims to automatically generate clinical reports from ear, nose, and throat (ENT) endoscopic images to assist clinical diagnosis. Unlike chest X-ray imaging, where a single image often provides relatively complete visual evidence, ENT endoscopy involves decomposed imaging, where multiple local views jointly support the overall diagnosis. Each view contains only partial anatomical and pathological information, while the clinical report summarizes findings from the entire examination. This discrepancy makes it challenging to derive holistic diagnostic semantics from distributed visual evidence and establish reliable correspondence between visual and textual representations. To address this challenge, we propose Holistic Semantic Learning for Endoscopic Report Generation (HSL-ERG), which learns holistic diagnostic semantics from distributed visual evidence under decomposed imaging. Specifically, the Holistic Semantic Representation Module (HSRM) aggregates distributed local visual evidence according to learned view contributions to construct a holistic examination-level semantic representation. Furthermore, the Reliable Correspondence Alignment Loss (RCAL) constructs adaptive bidirectional contrastive constraints between visual and textual representations to improve their correspondence reliability. Experiments on ENT-Endo, OCASD, and CTRG-Brain demonstrate the effectiveness of HSL-ERG. On ENT-Endo, HSL-ERG outperforms existing methods by 0.010 in BLEU-4 and 0.020 in clinical F1.

1 INTRODUCTION

Endoscopic Report Generation (ERG) aims to automatically generate clinical reports from endoscopic images acquired during examinations, thereby assisting clinical diagnosis. With recent advances in medical vision–language learning and multimodal generation, automatic report generation has been increasingly explored across various medical imaging scenarios Yao et al. (2025); Zhang et al. (2026); Kaklamanos et al. (2025). In Ear, Nose, and Throat (ENT) clinical practice, endoscopy is widely used to observe local anatomical structures, identify abnormal visual findings, and provide crucial visual evidence for diagnosis Qodirjonov (2026); Sakkizboyev & Makhmudov (2026). Automatic ERG can significantly reduce repetitive documentation, improve reporting efficiency, and facilitate intelligent clinical workflows.

However, ENT endoscopy differs substantially from conventional medical report generation in how diagnostic evidence is organized. As illustrated in Figure 1(a), an ENT examination typically contains multiple frames (averaging 6.54 frames per case), spanning multiple distinct or mixed anatomical regions (e.g., ear, nasal/sinus, and throat, as shown in Figure 1(b)). This characteristic defines a typical *decomposed imaging paradigm*. In radiological imaging (e.g., chest X-rays) or single-view diagnosis (Figure 1(c)), a single image often provides relatively complete visual evidence, yielding a direct global correspondence with the clinical report. In contrast, ENT endoscopy decomposes the examination into multiple non-registered local views (Figure 1(d)). Each view contains only partial anatomical and pathological information, while different views vary in diagnostic contribution, quality, or even suffer from missing information. Consequently, overall diagnostic semantics must be inferred from distributed, heterogeneous, and incomplete local evidence, making the visual–textual

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.

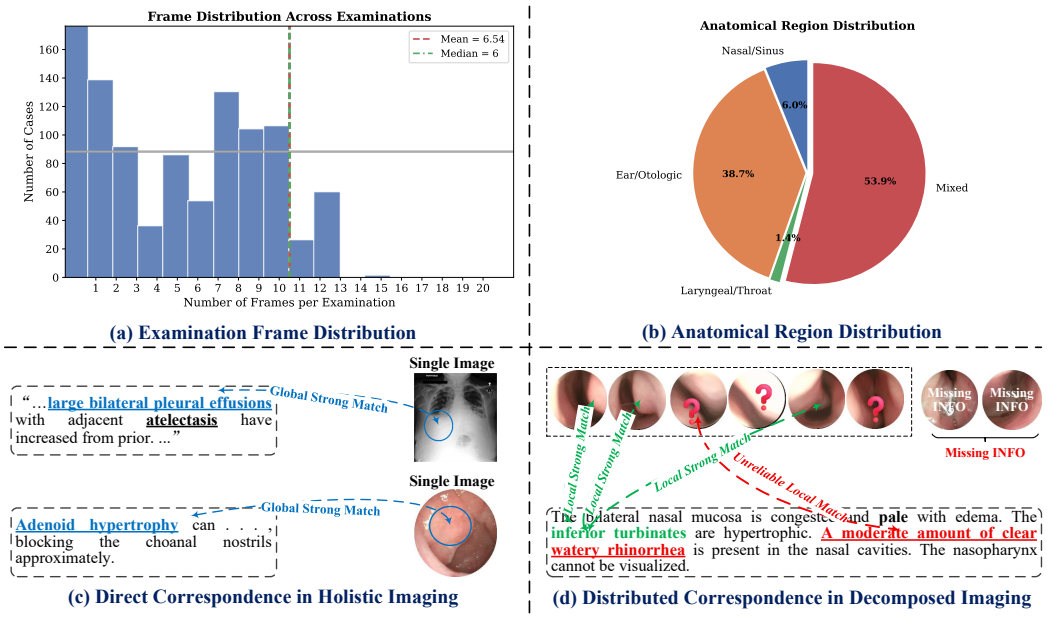


Figure 1: Motivation analysis and paradigm comparison. (a) Frame distribution per examination. (b) Anatomical region distribution. (c) Holistic imaging: A single image provides complete evidence for direct alignment. (d) Decomposed imaging: Distributed local views with partial evidence or missing information complicate overall semantic alignment.

correspondence highly unreliable and prone to misalignment. Existing endoscopic report generation methods mostly rely on straightforward feature aggregation and global cross-modal fusion Zhang et al. (2023); Naito et al. (2025); Kaklamanos et al. (2025). However, these methods overlook unequal diagnostic contributions across decomposed views and struggle to model the complex, indirect mapping between distributed local evidence and complete report semantics. Directly aggregating visual features easily introduces redundant information, which dilutes visual representation accuracy, while coarse global alignment fails to capture adaptive visual–textual correspondence under decomposed imaging. Therefore, learning a holistic examination-level semantic representation from distributed local views and establishing reliable visual–textual alignment remain fundamental challenges in ENT endoscopic report generation.

To address these challenges, we propose Holistic Semantic Learning for Endoscopic Report Generation (HSL-ERG). Specifically, the Holistic Semantic Representation Module (HSRM) aggregates distributed local visual evidence according to dynamically learned view contributions to construct a holistic, examination-level semantic representation. Furthermore, the Reliable Correspondence Alignment Loss (RCAL) enforces adaptive bidirectional contrastive constraints between visual and textual representations, improving correspondence reliability against local ambiguity and missing information. The main contributions of this work are summarized as follows:

- We formalize the correspondence challenge under the decomposed imaging paradigm in ENT endoscopy, highlighting how distributed, partial, and incomplete visual evidence hinders reliable global diagnostic alignment.
- We propose HSL-ERG, which integrates HSRM to dynamically aggregate local visual evidence based on view-specific diagnostic contributions, and RCAL to enforce reliable bidirectional cross-modal alignment.
- Extensive experiments on ENT-Endo, OCASD, and CTRG-Brain datasets demonstrate the effectiveness of HSL-ERG. On ENT-Endo, HSL-ERG outperforms existing methods by 0.010 in BLEU-4 and 0.020 in clinical F1 score.

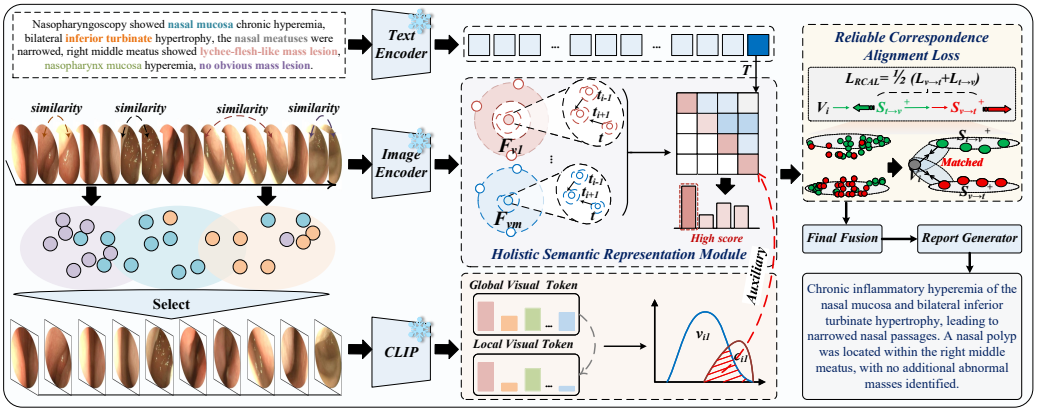


Figure 2: Overview of the HSL-ERG method. The Holistic Semantic Representation Module (HSRM) constructs an examination-level representation by modeling view contributions, while the Reliable Correspondence Alignment Loss (RCAL) enforces adaptive alignment for robust visual–textual correspondence.

2 Related Work

2.1 Endoscopic Report Generation

Endoscopic Report Generation (ERG) automatically generates clinical reports from endoscopic image sequences. Adapted from Radiology Report Generation (RRG) Kumar et al. (2026); Xiao et al. (2025); Mamdouh et al. (2025), ERG has evolved from CNN–RNN to Transformer-based, vision-language, and LLM architectures. While RRG methods leverage multi-slice CT/MRI Shu et al. (2026); Di Piazza et al. (2025); Caglayan et al. (2026); Park et al. (2026) or multi-view X-rays Chen et al. (2020; 2021) with stable spatial structures, ENT endoscopy relies on a decomposed paradigm where heterogeneous local views jointly form the diagnosis. However, existing ERG methods focus primarily on general visual-text generation, lacking specific modeling of the correspondence between distributed visual evidence and global diagnostic semantics.

2.2 Multi-scene Endoscopic Representation Learning

Intelligent endoscopic analysis has expanded from single-image tasks to multi-scene and multi-lesion representation learning Delgado et al. (2025); Kusters et al. (2025); Liu et al. (2026). Recent works employ topic-aware modeling, scene interaction, CNN–Transformer hybrids, mixture-of-experts, and multi-scale fusion to integrate multi-region information Nguyen et al. (2025c;a;b). Despite improving cross-image visual aggregation, these approaches treat local views as general visual features rather than complementary diagnostic evidence, leaving the alignment of local evidence with holistic diagnostic semantics underexplored.

2.3 Medical Image–Text Correspondence Learning

Medical vision-language learning aims to align medical images with clinical text for report generation Wan et al. (2023); Wang et al. (2024); Lai et al. (2026); Ryu et al. (2025); Shen et al. (2026); Algumaei et al. (2026). Recent works enhance cross-modal representation via medical knowledge integration, global-local alignment, and robust optimization Wang et al. (2022); Qin et al. (2022). However, current strategies optimize image-report or scene-text pairs as holistic units, which struggle in decomposed ENT endoscopy where fine-grained, distributed local views collectively support global diagnostic descriptions. Establishing reliable correspondence across such multi-view evidence remains a key challenge.

3 METHOD

3.1 OVERVIEW

As illustrated in Figure 2, HSL-ERG processes all available images from an examination to generate a comprehensive clinical report. To address the challenges of decomposed imaging, HSRM first evaluates view-specific relevance and integrates distributed local visual evidence into a holistic, examination-level semantic representation. Subsequently, RCAL enforces adaptive cross-modal contrastive constraints to establish reliable visual–textual alignment. Finally, the holistic representation guides a text decoder to generate the clinical report.

3.2 HOLISTIC SEMANTIC REPRESENTATION MODULE

ENT endoscopy typically involves decomposed imaging across multiple anatomical regions. Directly aggregating these views treats them equally, diluting key diagnostic evidence with redundant or weakly informative features. To address this, we propose the Holistic Semantic Representation Module (HSRM), which learns a holistic examination-level representation via semantic-enhanced local representation learning, local evidence weighting, and view contribution learning.

Given an input endoscopic image sequence $X = \{x_1, x_2, \ldots, x_N\}$ containing N images per examination, a primary visual encoder $f_v(\cdot)$ first extracts global visual features v_i^g and local visual tokens $V_i^l = \{v_{i1}, v_{i2}, \ldots, v_{iM}\}$ from each image x_i: $v_i^g, V_i^l = f_v(x_i)$, where M denotes the number of local visual tokens per image.

While primary local visual tokens mainly capture fine-grained structural and appearance details, they often lack high-level semantic context. Therefore, a pre-trained, frozen CLIP image encoder $f_c(\cdot)$ is introduced without fine-tuning to extract complementary semantic visual representations $c_i = f_c(x_i)$. Since features from the two encoders reside in different latent spaces, learnable linear projection matrices W_v and W_c align them into a shared embedding dimension:

$$\bar{v}_{ij} = W_v v_{ij}, \quad \bar{c}_i = W_c c_i. \tag{1}$$

The semantic-enhanced local representation $\hat{v}_{ij}$ is then obtained by injecting the frozen CLIP semantic prior into each local token via element-wise addition: $\hat{v}_{ij} = \bar{v}_{ij} + \bar{c}_i$.

In ENT endoscopic examinations, local region contributions vary significantly across anatomical sites. HSRM estimates the diagnostic importance weight a_{ij} for each local token using a parameter-efficient multi-layer perceptron (MLP) $\phi(\cdot)$:

$$a_{ij} = \frac{\exp\left(\phi(\hat{v}_{ij})\right)}{\sum_{k=1}^{M} \exp\left(\phi(\hat{v}_{ik})\right)}. \tag{2}$$

The aggregated image-level visual representation f_i^{vis} is then constructed by combining the weighted local features with the global feature: $f_i^{vis} = v_i^g + \sum_{j=1}^{M} a_{ij}\hat{v}_{ij}$.

Because individual frames in an examination exhibit varying diagnostic utility, HSRM estimates a view contribution weight α_i for each frame based on its image-level representation f_i^{vis}:

$$\alpha_i = \frac{\exp\left(\psi(f_i^{vis})\right)}{\sum_{k=1}^{N} \exp\left(\psi(f_k^{vis})\right)}, \tag{3}$$

where $\psi(\cdot)$ denotes a view contribution prediction network, α_i measures the relative diagnostic relevance of each view within the examination, adaptively suppressing the influence of redundant or weakly informative views. Finally, the examination-level holistic representation $F^{vis} = \sum_{i=1}^{N} \alpha_i f_i^{vis}$ is constructed via weighted aggregation, providing a unified visual representation to guide downstream clinical report generation.

3.3 RELIABLE CORRESPONDENCE ALIGNMENT LOSS

Although HSRM constructs a holistic examination-level visual representation, visually similar examinations across different patients may share overlapping clinical findings under the decomposed

imaging approach. Standard contrastive learning treats all unmatched image-report pairs within a mini-batch as strictly negative samples with uniform weights, pushing semantically related cases apart and impairing cross-modal alignment. To address this issue, we propose the Reliable Correspondence Alignment Loss (RCAL), which estimates examination-level visual similarity to adaptively suppress potential false-negative pairs during contrastive optimization.

For a mini-batch of B paired examination-report samples, HSRM first extracts examination-level visual representations $\mathcal{V} = \{F_1^{vis}, F_2^{vis}, \ldots, F_B^{vis}\}$. Concurrently, the associated clinical reports are processed by a text encoder $f_t(\cdot)$ to obtain textual representations $F_i^{txt} = f_t(T_i)$, forming the textual representation set $\mathcal{T} = \{F_1^{txt}, F_2^{txt}, \ldots, F_B^{txt}\}$. The cross-modal visual-textual similarity matrix $S \in \mathbb{R}^{B\times B}$ is computed via cosine similarity:

$$S_{ij} = \frac{F_i^{vis} \cdot F_j^{txt}}{\|F_i^{vis}\|\|F_j^{txt}\|}, \tag{4}$$

where S_{ii} denotes the true positive image-report pair, and S_{ij} $(i \neq j)$ denotes an unmatched pair.

Because examinations with similar visual characteristics frequently exhibit related clinical diagnoses, they are prone to acting as potential false negatives. To account for this, RCAL measures intra-modal visual similarity R_{ij} between different examinations:

$$R_{ij} = \frac{F_i^{vis} \cdot F_j^{vis}}{\|F_i^{vis}\|\|F_j^{vis}\|}. \tag{5}$$

Based on this similarity, an adaptive negative weighting factor w_{ij} is formulated to modulate the contribution of negative pairs:

$$w_{ij} = (B-1)\frac{\exp(-R_{ij}/\tau_r)}{\sum_{k\neq i}\exp(-R_{ik}/\tau_r)}, \quad \text{for } j \neq i, \tag{6}$$

where τ_r is a temperature parameter controlling weight sharpness. During gradient backpropagation, w_{ij} is treated as a detached scalar coefficient to prevent higher-order gradient interference and ensure stable optimization. The scaling factor $(B-1)$ preserves the total negative contrastive mass, satisfying $\sum_{j\neq i} w_{ij} = B-1$. Under uniform weighting (i.e., when visual similarities are isotropic), $w_{ij} = 1$, allowing RCAL to naturally reduce to standard InfoNCE. Higher visual similarity yields a smaller weight w_{ij}, thereby suppressing false-negative gradients during optimization.

Accordingly, the visual-to-text alignment loss $\mathcal{L}_{v\to t}$ and text-to-visual alignment loss $\mathcal{L}_{t\to v}$ are defined as:

$$\mathcal{L}_{v\to t} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(S_{ii}/\tau)}{\exp(S_{ii}/\tau) + \sum_{j\neq i} w_{ij}\exp(S_{ij}/\tau)}, \tag{7}$$

$$\mathcal{L}_{t\to v} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(S_{ii}/\tau)}{\exp(S_{ii}/\tau) + \sum_{j\neq i} w_{ij}\exp(S_{ji}/\tau)}, \tag{8}$$

where τ is the contrastive temperature parameter. Finally, the overall RCAL objective is defined as:

$$\mathcal{L}_{RCAL} = \frac{1}{2}(\mathcal{L}_{v\to t} + \mathcal{L}_{t\to v}). \tag{9}$$

3.4 Clinical Report Generation

Based on the examination-level visual representation F^{vis} learned by HSRM, a Transformer decoder is employed to autoregressively generate the clinical report. At the t-th decoding step, the generation probability is conditioned on the previously generated words $y_{<t}$ and the visual representation F^{vis}, formulated as $P(y_t \mid y_{<t}, F^{vis})$. The generation process is optimized using the standard cross-entropy loss $\mathcal{L}_{gen}$.

During training, the report generation objective and the Reliable Correspondence Alignment Loss are jointly optimized. The overall objective is formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{gen} + \lambda\mathcal{L}_{RCAL}, \tag{10}$$

where λ is a balancing hyperparameter that controls the trade-off between report generation and cross-modal alignment.

Table 1: Performance comparison across three input settings: Multi-view (ENT-Endo), Multi-slice (CTRG-Brain), and Single-image (OCASD). ENT-Endo utilizes all views per examination. Best and second-best results per dataset are marked in bold and underline, respectively.

Method	Venue	NLG Metrics ↑							CE Metrics ↑		
		B-1	B-2	B-3	B-4	MET	R-L	CID	P	R	F1
OCASD Dataset — Single-image											
R2Gen Chen et al. (2020)	EMNLP'20	0.508	0.369	0.283	0.222	0.218	0.455	1.045	**0.600**	0.541	0.569
R2GenCMN Chen et al. (2021)	ACL'21	0.524	0.368	0.274	0.212	0.226	0.431	1.020	0.534	0.544	0.539
TIMER Wu et al. (2023)	CHIL'23	0.533	0.392	0.304	0.242	0.226	0.433	1.289	0.561	0.549	0.555
R2GenGPT Wang et al. (2023)	Meta-Radiology'23	0.530	0.392	0.305	0.246	–	0.453	1.091	0.529	0.560	0.544
PromptMRG Jin et al. (2024)	AAAI'24	0.212	0.074	–	–	–	0.172	–	0.162	0.264	0.201
Medrat Hirsch et al. (2024)	ECCV'24	0.421	0.289	0.211	0.161	0.173	0.392	–	0.532	0.526	0.529
CMCRL Chen et al. (2025)	TIP'25	0.338	0.228	0.164	0.120	0.130	0.294	0.124	0.437	0.333	0.378
R2GenCSR Wang et al. (2026)	JBHI'26	0.461	0.325	0.244	0.189	**0.252**	0.378	0.869	0.475	0.469	0.472
HSL-ERG	Ours	**0.541**	**0.399**	**0.311**	**0.247**	0.226	**0.464**	**1.346**	0.590	**0.561**	**0.575**
CTRG-Brain Dataset — Multi-slice											
R2Gen-CMN Chen et al. (2021)	ACL'21	0.491	0.400	0.344	0.301	0.299	0.486	0.842	–	–	0.698
PGCA Shi et al. (2023)	EMNLP'24	0.502	0.378	0.307	0.265	0.325	0.430	0.340	–	–	0.692
PromptMRG Jin et al. (2024)	AAAI'24	0.481	0.383	0.316	0.265	0.310	0.474	0.503	–	–	0.692
HILT Liu et al. (2024)	arXiv'24	0.507	0.399	0.331	0.279	0.308	0.461	0.437	–	–	0.684
MEPNet Zhang et al. (2025)	AAAI'25	–	–	–	0.344	0.346	0.504	0.866	–	–	–
HSL-ERG	Ours	**0.519**	**0.403**	**0.351**	**0.349**	**0.347**	**0.508**	**0.871**	–	–	**0.709**
ENT-Endo Dataset — Multi-view											
R2Gen Chen et al. (2020)	EMNLP'20	0.463	0.336	0.252	0.196	0.202	0.412	0.637	0.550	0.490	0.518
R2GenCMN Chen et al. (2021)	ACL'21	0.496	0.355	0.266	0.207	0.203	0.401	0.632	0.508	0.500	0.504
TIMER Wu et al. (2023)	CHIL'23	0.502	0.360	0.267	0.208	0.213	0.399	0.625	0.513	0.515	0.514
R2GenGPT Wang et al. (2023)	Meta-Radiology'23	0.476	0.360	0.278	0.224	–	0.412	0.668	–	–	–
PromptMRG Jin et al. (2024)	AAAI'24	0.385	0.280	0.205	0.155	0.218	0.343	0.239	0.456	0.507	0.480
Medrat Hirsch et al. (2024)	ECCV'24	0.419	0.290	0.217	0.169	0.171	0.380	–	0.477	0.433	0.454
CMCRL Chen et al. (2025)	TIP'25	0.285	0.177	0.121	0.088	0.125	0.291	0.157	0.450	0.326	0.378
R2GenCSR Wang et al. (2026)	JBHI'26	0.450	0.328	0.249	0.194	**0.240**	0.364	0.698	0.471	0.444	0.457
HSL-ERG	Ours	**0.515**	**0.390**	**0.299**	**0.234**	0.223	**0.428**	**0.709**	**0.559**	**0.518**	**0.538**

4 EXPERIMENTS

4.1 DATASETS

ENT-Endo contains 1,106 ENT endoscopic examinations with 7,232 frames and expert-reviewed case-level reports. Each examination contains multiple local views (6.54 frames on average), which jointly support the diagnosis; we therefore classify it as *Multi-view*. The data Liu et al. (2025) are split 70%/10%/20% for training/validation/test. **CTRG-Brain** Tang et al. (2024) contains 6,001 brain CT cases with 160,352 slices and 6,001 clinical reports. After preprocessing Liu et al. (2025), each case is represented by 24 slices and split 7:2:1 into training/validation/test. It is used as an independent *Multi-slice* validation setting rather than as an ENT-specific dataset. **OCASD** Viscaino et al. (2020); Khan & Khan (2024) contains 500 expert-annotated otoscopic images across five categories, each paired with an expert-written summary, and is split 70%/10%/20% Liu et al. (2025); Zou et al. (2025). It is a *Single-image* setting and therefore provides complementary endoscopic validation rather than direct decomposed multi-view validation.

4.2 EVALUATION METRICS

We evaluate the generated reports using both Natural Language Generation (NLG) metrics and Clinical Efficacy (CE) metrics to assess linguistic quality and clinical consistency. The NLG metrics include BLEU Papineni et al. (2002), METEOR Banerjee & Lavie (2005), ROUGE-L Lin (2004), and CIDEr Vedantam et al. (2015), which evaluate the linguistic quality and content consistency between the generated reports and the reference reports. For the nasal endoscopy dataset, the CE metrics are evaluated based on the pathological finding categories annotated in the dataset. For the otoscopic dataset, the CE metrics are evaluated according to the disease diagnosis categories defined

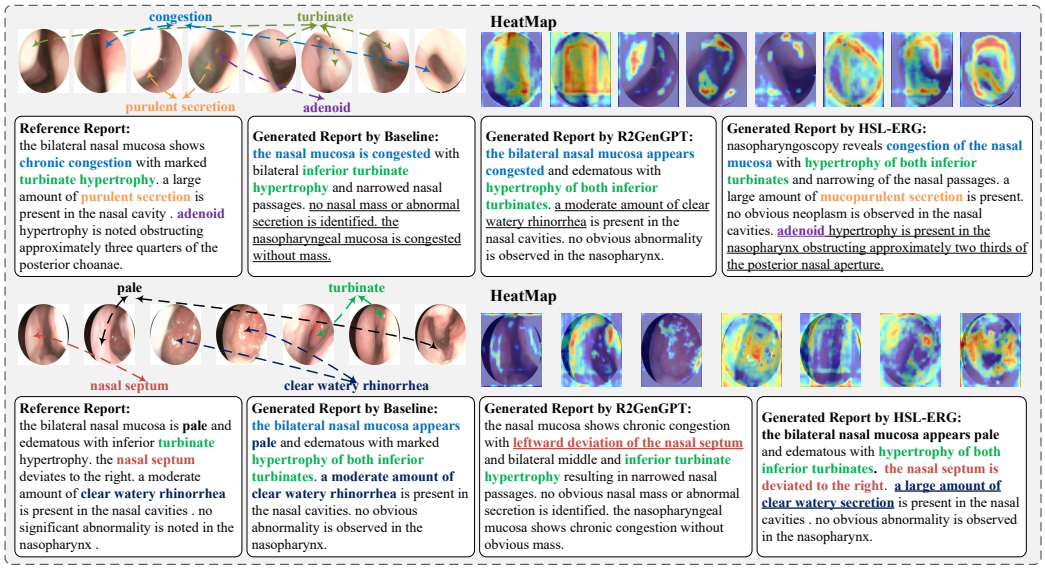


Figure 3: Qualitative comparison on ENT-Endo cases. HSL-ERG produces more accurate reports with consistent focus on diagnostic regions.

in the dataset. Precision, Recall, and F1-score are adopted to evaluate the category-level consistency between the predicted results and the reference annotations.

4.3 Implementation Details

Models are implemented in PyTorch and trained on a single NVIDIA RTX 3090 GPU (24 GB VRAM) with input images resized to 224×224. For feature extraction, an ImageNet-pretrained ResNet-101 serves as the primary visual encoder $f_v(\cdot)$, complemented by a pre-trained CLIP (ViT-B/16) as the semantic encoder $f_c(\cdot)$. Notably, the CLIP encoder remains entirely frozen to preserve its generalizable semantic priors and prevent over-fitting on small medical datasets; domain adaptation is mediated via learnable projection matrices (W_c) and the primary encoder. Optimization uses Adam with weight decay 5×10^{-5}. The 5×10^{-5} learning rate applies exclusively to the primary visual encoder (ResNet-101), while the encoder-decoder module is trained at 1×10^{-4}, scheduled by StepLR ($\gamma = 0.1$). Training lasts for 50 epochs on ENT-Endo (batch size = 4) and 10 epochs on CTRG-Brain. For RCAL, $\tau_r = 0.02$ and $\lambda = 0.8$.

4.4 Comparison with Existing Methods

Table 1 compares HSL-ERG with representative report generation methods across three clinical input settings: multi-view (ENT-Endo), multi-slice (CTRG-Brain), and single-image (OCASD). Because specialized ENT endoscopic report generation methods remain scarce, we evaluate approaches across different paradigms under uniform conditions.

On the primary ENT-Endo benchmark, HSL-ERG outperforms competing baselines on nearly all metrics, achieving 0.234 BLEU-4, 0.428 ROUGE-L, 0.709 CIDEr, and 0.538 F1 score. Compared with the strongest competing results, HSL-ERG improves BLEU-4 from 0.224 to 0.234, CIDEr from 0.698 to 0.709, and clinical entity F1 from 0.518 to 0.538. These gains demonstrate that HSL-ERG better preserves critical diagnostic evidence distributed across local views. This validates our design logic: HSRM aggregates distributed evidence via learned view contributions, while RCAL tightens cross-modal alignment between examination-level visual features and report semantics.

On the multi-slice CTRG-Brain dataset, HSL-ERG achieves state-of-the-art performance across all metrics, reaching 0.349 BLEU-4, 0.508 ROUGE-L, 0.871 CIDEr, and 0.709 F1 score. Compared to the second-best results, it advances BLEU-4 from 0.344 to 0.349, CIDEr from 0.866 to 0.871, and F1 from 0.698 to 0.709, demonstrating strong generalizability to 3D volumetric scenarios.

Table 2: Sensitivity to the input-view budget on ENT-Endo. Avg. Views denotes the average number of valid views used.

Budget	Avg.	NLG		CE		
		B-4	CID	P	R	F1
1	1.0	0.187	0.537	0.542	0.479	0.508
2	2.0	0.200	0.601	0.514	0.521	0.518
4	4.0	0.192	0.600	0.561	0.491	0.524
6	6.0	0.209	0.685	0.579	0.488	0.530
All	6.5	0.234	0.709	0.559	0.518	0.538

Table 3: Effect of RCAL weight λ on the ENT-Endo dataset under a fixed 20-epoch training schedule.

λ	B-4	MET	R-L	CID	P	R	F1
0.5	0.210	0.211	0.422	0.511	0.555	0.504	0.528
0.8	**0.215**	**0.212**	**0.424**	**0.593**	**0.556**	**0.508**	**0.531**
1.0	0.199	0.197	0.411	0.552	0.542	0.499	0.519
2.0	0.187	0.190	0.408	0.483	0.520	0.493	0.506

Table 4: Ablation study of the proposed HSRM and RCAL on the ENT-Endo and OCASD datasets. ✓ indicates that the corresponding module is enabled. The best and second-best results are highlighted in bold and underline, respectively.

ID	Modules		ENT-Endo						OCASD					
	HSRM	RCAL	B-4	MET	R-L	P	R	F1	B-4	MET	R-L	P	R	F1
A0			0.211	0.210	0.410	0.520	0.495	0.507	0.222	0.218	0.455	**0.600**	0.541	0.569
A1	✓		<u>0.220</u>	0.208	**0.432**	**0.562**	<u>0.505</u>	<u>0.532</u>	<u>0.245</u>	<u>0.223</u>	<u>0.462</u>	<u>0.597</u>	<u>0.551</u>	<u>0.573</u>
A2		✓	0.215	<u>0.211</u>	0.422	0.557	0.502	0.528	0.241	<u>0.223</u>	0.460	0.587	0.539	0.562
A3	✓	✓	**0.234**	**0.223**	<u>0.428</u>	<u>0.559</u>	**0.518**	**0.538**	**0.247**	**0.226**	**0.464**	0.590	**0.561**	**0.575**

On the single-image OCASD dataset, HSL-ERG similarly leads the benchmark with 0.247 BLEU-4, 0.464 ROUGE-L, 1.346 CIDEr, and 0.575 F1 score, advancing CIDEr from 1.289 to 1.346 and F1 from 0.569 to 0.575 over the second-best results. As a single-image baseline, OCASD verifies that our cross-modal alignment mechanism remains highly effective even without multi-view aggregation.

4.5 Ablation Study

4.5.1 Effectiveness of Core Modules

Table 4 evaluates the individual and complementary contributions of HSRM and RCAL in mitigating representation ambiguity and cross-modal misalignment under decomposed endoscopic imaging. On the primary ENT-Endo dataset where diagnostic evidence is distributed across heterogeneous local views, enabling HSRM alone (A1) increases BLEU-4 from 0.211 to 0.220 and F1 from 0.507 to 0.532 over the baseline (A0), indicating that reliability-aware aggregation preserves complementary visual evidence. Enabling RCAL alone (A2) raises BLEU-4 and F1 to 0.215 and 0.528, respectively, showing that adaptive alignment tightens the correspondence between examination-level visual representations and overall report semantics. Combining both modules (A3) achieves the best performance, reaching 0.234 BLEU-4, 0.518 Recall, and 0.538 F1 score. Compared with A0, the complete framework yields net improvements of 0.023 in BLEU-4 and 0.031 in F1 score, confirming the synergism between reliable evidence aggregation and correspondence optimization. On the single-image OCASD dataset, the complete model similarly attains top performance in BLEU-4, METEOR, ROUGE-L, Recall, and F1 score, validating its versatility across different endoscopic configurations.

4.5.2 Sensitivity to the Number of Input Views

As shown in Table 2, we evaluate model performance under view budgets $K \in \{1, 2, 4, 6, \text{All}\}$. For an examination containing N_i images, the model processes $\min(N_i, K)$ valid views, whereas *All* retains all available images. When $N_i > K$, views are randomly sampled during training and fixed during evaluation; otherwise, all valid views are retained and padded positions are masked. Each setting is trained independently. Increasing the view budget from 1 to All steadily improves BLEU-4 from 0.187 to 0.234, CIDEr from 0.537 to 0.709, and F1 score from 0.508 to 0.538. This performance trajectory confirms that ENT endoscopic diagnosis depends heavily on complementary multi-view evidence, and demonstrates that HSRM effectively scales with expanding view counts without accumulating redundant visual information.

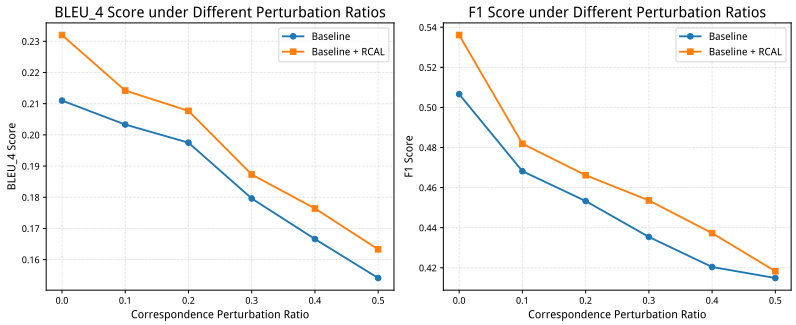


Figure 4: Performance comparison under different correspondence perturbation ratios on the ENT-Endo dataset.

4.5.3 Analysis of RCAL Weight λ

Table 3 evaluates different RCAL loss weights on ENT-Endo under a fixed 20-epoch schedule. The optimal performance is obtained at $\lambda = 0.8$, yielding 0.215 BLEU-4, 0.593 CIDEr, and 0.531 F1 score. Increasing λ to 2.0 reduces BLEU-4 and CIDEr to 0.187 and 0.483, respectively, suggesting that excessive alignment constraints may compete with primary report generation. We thus fix $\lambda = 0.8$ for final model training.

4.5.4 Robustness under Correspondence Perturbations

Figure 4 evaluates the baseline and RCAL under increasing correspondence perturbation ratios, where a proportion of training report labels are randomly swapped while keeping the test set intact. Although performance declines for both models under higher perturbation ratios, RCAL consistently maintains superior BLEU-4 and F1 scores with a significantly milder degradation. These results demonstrate that RCAL, through its adaptive weighting mechanism, effectively improves cross-modal alignment robustness against image-report pairing uncertainties.

4.6 Qualitative Analysis

Figure 3 compares reports generated by different models against the reference standard. The Baseline often omits key findings, while R2GenGPT exhibits clinical detail inconsistencies despite fluent text. In contrast, our model accurately captures complementary findings—including mucosal congestion, inferior turbinate hypertrophy, nasal secretion, and adenoid hypertrophy—aligning closely with the reference. Attention maps confirm consistent focus on diagnostic regions, validating multi-view evidence aggregation under decomposed settings. Detailed case analysis reveals that while core categories are correctly identified, minor discrepancies persist in fine-grained attributes like severity grading (e.g., “moderate” vs. “copious”) and numerical estimation (e.g., “three-quarters” vs. “two-thirds”), showing that precise continuous descriptor modeling remains challenging.

5 Conclusion

In this paper, we proposed HSL-ERG for learning holistic diagnostic semantics from distributed evidence under decomposed imaging. HSRM aggregates local evidence according to learned view contributions, while RCAL improves visual–text correspondence by downweighting potential false negatives. Experiments across ENT-Endo, CTRG-Brain, and OCASD validate the effectiveness of our semantic aggregation and alignment mechanisms. Nevertheless, standard metrics still lack sensitivity to fine-grained clinical variations in severity, localization, and numerical estimation. Future work will explore granular clinical truthfulness evaluation protocols and attribute-level alignment constraints to further enhance clinical dependability.

AI USE STATEMENT

Generative AI tools were used during the preparation of this paper solely to assist with text drafting, language polishing, and LaTeX formatting. All content was thoroughly reviewed and verified by the authors, who assume full responsibility for the accuracy and integrity of the manuscript.

REFERENCES

Ali Algumaei, Wassim Bouachir, and Nizar Bouguila. Medical vision-language models for robust disease diagnosis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6104–6113, 2026.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Lara Caglayan, Timur Ugurlu, Nicolae Bolog, Daehyun Yoon, Younèss Nour, Anne Péporté, Gustav Andreisek, and Cornelia Brendle. Structured vs. unstructured mri reporting for lumbar spine degenerative conditions: a comparative evaluation using a custom radiology template. *Skeletal Radiology*, 55(1):81–95, 2026.

Weixing Chen, Yang Liu, Ce Wang, Jiarui Zhu, Guanbin Li, Cheng-Lin Liu, and Liang Lin. Cross-modal causal representation learning for radiology report generation. *IEEE Transactions on Image Processing*, 2025.

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 (volume 1: long papers)*, pp. 5904–5914, 2021.

Lucas Monteiro Delgado, Gilmara Coelho Meine, Paula Santo, Caio Mendonça Magalhães, Vitor Lauar Pimenta de Figueiredo, Josue Aliaga Ramos, and Vitor Nunes Arantes. Endoscopic submucosal tunneling techniques versus flexible endoscopic septotomy for zenker's diverticulum: a systematic review and meta-analysis. *Gastrointestinal Endoscopy*, 101(4):751–761, 2025.

Theo Di Piazza, Carole Lazarus, Olivier Nempont, and Loic Boussel. Ct-agrg: Automated abnormality-guided report generation from 3d chest ct volumes. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pp. 01–05. IEEE, 2025.

Elad Hirsch, Gefen Dawidowicz, and Ayellet Tal. Medrat: Unpaired medical report generation via auxiliary tasks. In *European Conference on Computer Vision*, pp. 18–35. Springer, 2024.

Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. Promptmrg: Diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 2607–2615, 2024.

Evandros Kaklamanos, Kristjana Kristinsdottir, Jonathan Huang, Dustin Carlson, Rajesh Keswani, John Pandolfino, and Mozziyar Etemadi. From scope to script: An automated report generation model for gastrointestinal endoscopy. In *MICCAI Workshop on Deep Generative Models*, pp. 131–140. Springer, 2025.

Eram Anwarul Khan and Anas Anwarul Haq Khan. Sumotosima: A framework and dataset for classifying and summarizing otoscopic images. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pp. 1–11, 2024.

Satyam Kumar, Kaustubh Shivshankar Shejole, and Pushpak Bhattacharyya. Looking at radiology report generation through a causal lens: A survey. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4029–4057, 2026.

Carolus HJ Kusters, Tim JM Jaspers, Tim GW Boers, Martijn R Jong, Jelmer B Jukema, Kiki N Fockens, Albert J De Groof, Jacques J Bergman, Fons Van Der Sommen, and Peter HN De With. Will transformers change gastrointestinal endoscopic image analysis? a comparative analysis between cnns and transformers, in terms of performance, robustness and generalization. *Medical Image Analysis*, 99:103348, 2025.

Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, Yuheng Li, Konstantinos Psounis, and Xiaofeng Yang. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *IEEE transactions on medical imaging*, 2026.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Che Liu, Zhongwei Wan, Yuqi Wang, Hui Shen, Haozhe Wang, Kangyu Zheng, Mi Zhang, and Rossella Arcucci. Benchmarking and boosting radiology report generation for 3d high-resolution medical images. *arXiv preprint arXiv:2406.07146*, 1(2), 2024.

Shengyuan Liu, Boyun Zheng, Wenting Chen, Zhihao Peng, Zhenfei Yin, Jing Shao, Jiancong Hu, and Yixuan Yuan. Endobench: A comprehensive evaluation of multi-modal large language models for endoscopy analysis. *Advances in Neural Information Processing Systems*, 38, 2026.

Tengfei Liu, Jiapu Wang, Yongli Hu, Mingjie Li, Junfei Yi, Xiaojun Chang, Junbin Gao, and Baocai Yin. Hc-llm: Historical-constrained large language models for radiology report generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 5595–5603, 2025.

Dima Mamdouh, Mariam Attia, Mohamed Osama, Nesma Mohamed, Abdelrahman Lotfy, Tamer Arafa, Essam A Rashed, and Ghada Khoriba. Advancements in radiology report generation: A comprehensive analysis. *Bioengineering*, 12(7):693, 2025.

Tatsushi Naito, Takuto Nosaka, Tomoko Tanaka, Yu Akazawa, Kazuto Takahashi, Masahiro Ohtani, and Yasunari Nakamoto. Usefulness of an artificial intelligence-based colonoscopy report generation support system. *Clinical endoscopy*, 58(2):327–330, 2025.

Trong-Nhan Nguyen, Luan LM Nguyen, Phat-Dat To, Tran-Quoc Duy Nguyen, Anh-Huy Nguyen, Tuan Pham-Dang, Chu Lam Nguyen, and Duy VM Nguyen. Hymoenet: Mixture-of-experts enhanced cnn-transformer hybrid framework for classifying anatomical sites in endoscopic ent images. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 14184–14189, 2025a.

Trong-Thuan Nguyen, Viet-Tham Huynh, Thao Thi Phuong Dao, Mai-Khiem Tran, Ha Nguyen Thi, Tien To Vu Thuy, Uyen Hanh Tran, Tam V Nguyen, Minh-Triet Tran, and Thanh Dinh Le. Acm multimedia grand challenge on ent endoscopy analysis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 14178–14183, 2025b.

Y Hop Nguyen, Doan Anh Phan Huu, Trung Thai Tran, Nhat Nam Mai, Van Toi Giap, Thao Thi Phuong Dao, and Trung-Nghia Le. Multi-level cls token fusion for contrastive learning in endoscopy image classification. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 14197–14203, 2025c.

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*, pp. 311–318, 2002.

Jiwoo Park, Kyunghwa Han, Ji Seon Oh, Hee Dong Chae, Ahram Kim, Si Young Park, Hye Jin Yoo, and Young Han Lee. Prospective evaluation of artificial intelligence (ai) in lumbar spine magnetic resonance imaging (mri) workflow: from deep learning (dl)-enhanced accelerated acquisition to simultaneous vision-language model (vlm)-based automated report generation. *European Journal of Radiology*, pp. 112695, 2026.

Yang Qin, Dezhong Peng, Xi Peng, Xu Wang, and Peng Hu. Deep evidential learning with noisy correspondence for cross-modal retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 4948–4956, 2022.

Ikromjon Qodirjonov. Artificial intelligence–driven radiomics in otorhinolaryngology: Enhancing diagnosis and outcome prediction in head and neck cancer. *International Journal of Clinical & Translational Medicine*, 1(2):1–6, 2026.

Ji Seung Ryu, Hyunyoung Kang, Yuseong Chu, and Sejung Yang. Vision-language foundation models for medical imaging: a review of current practices and innovations. *Biomedical Engineering Letters*, 15(5):809–830, 2025.

Ismoiljon Sakkizboyev and Uktamjon Makhmudov. Advances in the diagnosis and management of common otorhinolaryngological disorders: An updated narrative review. *Journal of Clinical and Biomedical Research*, 1(1):126–131, 2026.

Lingdong Shen, Xiaoshuang Huang, Fangxin Shang, Xudong Zhang, Yehui Yang, Bin Fan, and Shiming Xiang. From image to pixels: towards fine-grained medical vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Yanzhao Shi, Junzhong Ji, Xiaodan Zhang, Liangqiong Qu, and Ying Liu. Granularity matters: pathological graph-driven cross-modal alignment for brain ct report generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6617–6630, 2023.

Xiwen Shu, Ke Lu, Jingtong Zhao, John Ginn, Yongbok Kim, Zhenyu Yang, Justus Adamson, Trey Mullikin, and Chunhao Wang. Knowledge-based deep residual u-net (dru) for synthetic ct generation using a single mr volume for frameless radiosurgery. *Journal of Applied Clinical Medical Physics*, 27(1):e70343, 2026.

Yuhao Tang, Haichen Yang, Liyan Zhang, and Ye Yuan. Work like a doctor: Unifying scan localizer and dynamic generator for automated computed tomography report generation. *Expert Systems with Applications*, 237:121442, 2024.

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*, pp. 4566–4575, 2015.

Michelle Viscaino, Juan Cristóbal Maass Oñate, Paul Délano Reyes, Mariela Torrente Avendaño, Carlos Stott Caro, and Fernando Auat Cheein. Computer-aided diagnosis of external and middle ear conditions: A machine learning approach. 2020.

Zhongwei Wan, Che Liu, Mi Zhang, Jie Fu, Benyou Wang, Sibo Cheng, Lei Ma, César Quilodrán-Casas, and Rossella Arcucci. Med-unic: Unifying cross-lingual medical vision-language pre-training by diminishing bias. *Advances in Neural Information Processing Systems*, 36:56186–56197, 2023.

Pengyu Wang, Huaqi Zhang, and Yixuan Yuan. Mcpl: multi-modal collaborative prompt learning for medical vision-language model. *IEEE Transactions on Medical Imaging*, 43(12):4224–4235, 2024.

Xiao Wang, Yuehang Li, Fuling Wang, Shiao Wang, Chuanfu Li, and Bo Jiang. R2gencsr: Mining contextual and residual information for llms-based radiology report generation. *IEEE Journal of Biomedical and Health Informatics*, 2026.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. R2gengpt: Radiology report generation with frozen llms. *Meta-Radiology*, 1(3):100033, 2023.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3876–3887, 2022.

Yuexin Wu, I-Chan Huang, and Xiaolei Huang. Token imbalance adaptation for radiology report generation. In *Conference on Health, Inference, and Learning*, pp. 72–85. PMLR, 2023.

Ting Xiao, Lei Shi, Peng Liu, Zhe Wang, and Chenjia Bai. Radiology report generation via multi-objective preference optimization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 8664–8672, 2025.

Jieru Yao, Xueran Li, Qiang Xie, Longfei Han, Yiwen Jia, Nian Liu, Dingwen Zhang, and Junwei Han. Llava-endo: a large language-and-vision assistant for gastrointestinal endoscopy. *Frontiers of Computer Science*, 19(4):194331, 2025.

Lihui Zhang, Zihua Lu, Liwen Yao, Zehua Dong, Wei Zhou, Chunping He, Renquan Luo, Mengjiao Zhang, Jing Wang, Yanxia Li, et al. Effect of a deep learning–based automatic upper gi endoscopic reporting system: a randomized crossover study (with video). *Gastrointestinal Endoscopy*, 98(2): 181–190, 2023.

Xiaodan Zhang, Yanzhao Shi, Junzhong Ji, Chengxin Zheng, and Liangqiong Qu. Mepnet: Medical entity-balanced prompting network for brain ct report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25940–25948, 2025.

Xingyu Zhang, Xiangwei Zheng, Zhen Li, Mingjun Ma, Xiao Liang, Xiuxiu Ren, and Mingzhe Zhang. A fine-tuning multimodal large language model for endoscopic report generation. *Biomedical Signal Processing and Control*, 118:109737, 2026.

Ke Zou, Yang Bai, Bo Liu, Yidi Chen, Zhihao Chen, Yang Zhou, Xuedong Yuan, Meng Wang, Xiaojing Shen, Xiaochun Cao, et al. Uncertainty-aware medical diagnostic phrase identification and grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

A Appendix

A.1 Dataset Details

The ENT-Endo Dataset comprises 1,106 clinical endoscopic examinations with 7,232 RGB frames collected from routine ENT procedures. Each examination is paired with a free-text English clinical report at the case level, where multiple endoscopic frames jointly correspond to a single diagnostic description. This case-level annotation setting reflects the decomposed imaging paradigm of ENT endoscopy: images are captured independently from different local anatomical regions rather than forming a spatially continuous volume. Consequently, local views within an examination jointly provide the visual evidence required for overall clinical assessment. A representative case-level image set is illustrated in Figure 5.

The number of frames per examination varies substantially, with a mean of 6.54 frames, a median of 6 frames, and a range of 1–20 frames. In particular, 42.7% of examinations contain 6–10 frames, indicating that most examinations require multiple local observations to capture comprehensive clinical findings. The overall frame-number distribution and its consistency across training, validation, and test sets are presented in Figure 6a. Detailed statistics for the three subsets are summarized in Table 5.

The dataset covers three major anatomical categories: ear/otologic, nasal/sinus, and laryngeal regions. Among all examinations, 53.3% involve mixed anatomical regions, in which observations from multiple regions, such as the nasal cavity and ear, are described within a single case-level clinical report. However, reports do not provide explicit region-level correspondence between individual sentences and local images. This characteristic further reflects the complexity of decomposed endoscopic imaging and introduces ambiguity when establishing fine-grained image–text correspondence. The anatomical distribution of the dataset is summarized in Table 6.

The dataset is divided into training, validation, and test sets containing 774, 110, and 222 examinations, respectively. The three subsets maintain comparable frame-number distributions, thereby reducing potential distribution bias during model training and evaluation, as shown in Figure 6a.

Another notable characteristic of the ENT-Endo Dataset is the long-tailed distribution of clinical findings. Since annotations are provided as free-text reports rather than predefined diagnostic labels, each examination may contain multiple co-occurring findings. As shown in Figure 6b, the most frequent finding, mucosal abnormality, appears in 55.5% of examinations, whereas the relatively rare finding, polyp, occurs in only 1.4% of cases, resulting in an approximately 40-fold frequency gap. This imbalance reflects the natural distribution of findings in routine ENT clinical practice while introducing additional challenges for accurately generating descriptions of infrequent abnormalities.

Table 5: Overview of the ENT-Endo Dataset.

Split	Cases	Frames	Frames/Case (mean±std)
Train	774	5,101	6.59±3.77
Val	110	714	6.49±3.70
Test	222	1,417	6.38±3.84
Total	**1,106**	**7,232**	**6.54±3.78**

Table 6: Anatomical region distribution inferred from clinical reports.

Anatomical Region	Cases	Percentage (%)
Mixed (multiple regions)	590	53.3
Ear / Otologic	423	38.2
Nasal / Sinus	66	6.0
Laryngeal / Throat	15	1.4
Total	**1,106**	**100.0**

Overall, the ENT-Endo Dataset exhibits three important characteristics. First, each examination is represented by a variable-sized set of independently captured local images rather than a spatially continuous image volume. Second, multiple local views jointly correspond to a single case-level report without explicit image–sentence alignment. Third, clinical findings follow a markedly long-tailed distribution. These characteristics make the dataset suitable for evaluating case-level endoscopic report generation under decomposed imaging and weak image–text correspondence.



Original report (Patient 8208):
The bilateral nasal mucosa shows mild congestion. Both **inferior turbinates** are hypertrophied. The **nasal septum** is deviated. **Adenoid** hyperplasia and enlargement are observed in the nasopharynx, obstructing approximately two-thirds of the posterior choanae.

Figure 5: A representative case-level image set in the ENT-Endo Dataset. Multiple images capture local views from different anatomical regions and jointly correspond to a single clinical report.

A.2 Training Algorithm

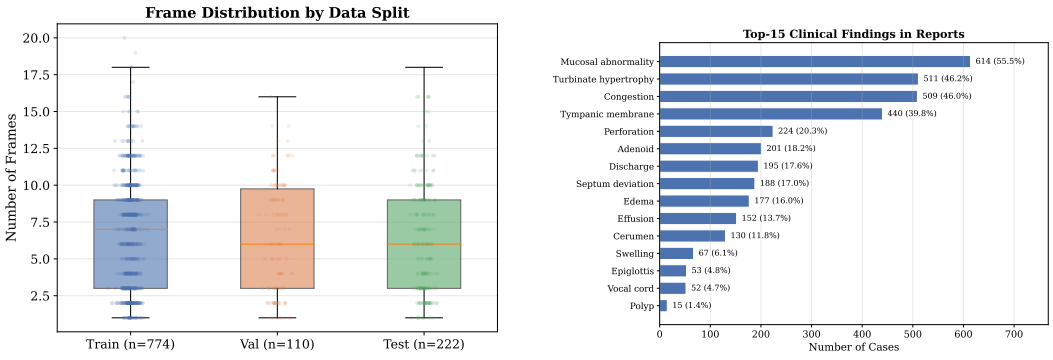
To facilitate reproducibility, we provide the overall training procedure of HSL-ERG in Algorithm 1. Given multiple endoscopic images from an examination, the visual encoder first extracts local feature maps, while HSRM models view-level contributions to aggregate examination-level representations. The cross-modal alignment objective RCAL is optimized to refine visual–text correspondence by downweighting potential false negatives. Finally, the aggregated representations are fed into the report decoder for clinical report generation.

Algorithm 1 Training Procedure of HSL-ERG

Require: Multi-view images $\mathcal{I} = \{I_1, \ldots, I_N\}$, clinical report Y, loss weight λ
Ensure: Generated report $\hat{Y}$
1: Extract local visual features via visual encoder: $\mathbf{V} = E_v(\mathcal{I})$
2: Extract frozen CLIP-based image and text features: $\mathbf{V}_c = E^v_{clip}(\mathcal{I})$, $\mathbf{T}_c = E^t_{clip}(Y)$
3: Compute holistic examination-level visual representation via HSRM: $\tilde{\mathbf{V}} = \mathrm{HSRM}(\mathbf{V}, \mathbf{V}_c)$
4: Compute Relational Correspondence Alignment Loss: $\mathcal{L}_{rcal} = \mathcal{L}_{RCAL}(\tilde{\mathbf{V}}, \mathbf{T}_c)$
5: Generate report via language decoder: $\hat{Y} = D(\tilde{\mathbf{V}})$
6: Compute cross-entropy generation loss: $\mathcal{L}_{gen} = \mathrm{CE}(\hat{Y}, Y)$
7: Optimize full model with joint loss: $\mathcal{L} = \mathcal{L}_{gen} + \lambda \mathcal{L}_{rcal}$
8: **return** $\hat{Y}$

A.3 Visualization of Multi-Frame Reliability Scoring

Figure 7 presents the multi-frame scoring results of two representative ENT endoscopic examinations. Each examination contains multiple local images captured from different anatomical regions



(a) Frame distribution across dataset splits

(b) Clinical finding frequency

Figure 6: Dataset statistics of the ENT-Endo dataset. (a) Frame distribution comparison among training, validation, and test sets. (b) Clinical finding frequency inferred from clinical reports.

or observation positions, while all images jointly correspond to a single case-level clinical report. For each case, the top row shows original endoscopic images, the middle heatmap presents pairwise frame-to-frame CLIP similarities, and the bottom bar chart reports the alignment score between each individual frame and the corresponding clinical report. The visualization shows that images within the same examination generally maintain relatively high semantic similarity. However, their alignment scores with the case-level report are clearly non-uniform. Some frames provide visual evidence strongly associated with reported clinical findings, whereas others mainly contain repeated observations, complementary anatomical information, or relatively weak report-related evidence. Therefore, high similarity among local images does not necessarily indicate that they contribute equally to the overall diagnosis. These results provide an intuitive motivation for HSRM. Instead of uniformly aggregating all local representations, HSRM models semantic relationships among different views and estimates their relative reliability. It consequently emphasizes informative and report-relevant visual evidence while reducing the influence of redundant or weakly aligned frames, thereby constructing a more reliable case-level visual representation for report generation.

A.4 Visualization of View-Count Sensitivity

Figure 8 visualizes the case-level representations learned under different input-view budgets using t-SNE. Green triangles denote ground-truth report representations, while colored markers correspond to visual representations obtained with $K \in \{1, 2, 4, 6, \mathrm{All}\}$. When only one input view is used, learned representations are widely dispersed, reflecting limited clinical evidence available from a single local image. As the number of input views increases, representations form more compact and semantically structured distributions. This qualitative pattern indicates that integrating multiple local views provides complementary information for constructing a more complete case-level representation. Representations obtained with $K = 6$ and All exhibit broadly similar spatial organization, suggesting that a moderate number of complementary views can capture most clinically relevant evidence, whereas retaining all available images may provide limited additional information.

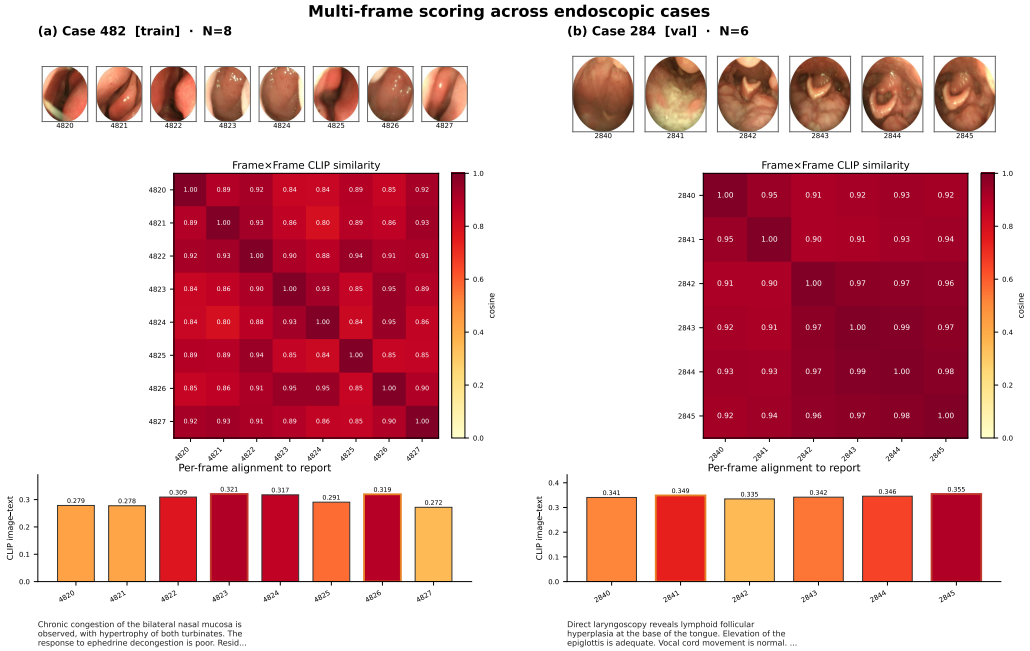


Figure 7: Multi-frame scoring visualization in HSRM for two ENT cases. Top: local images; Middle: pairwise frame similarities; Bottom: per-frame alignment scores with the report. Non-uniform scores highlight varying view contributions to case-level representations.

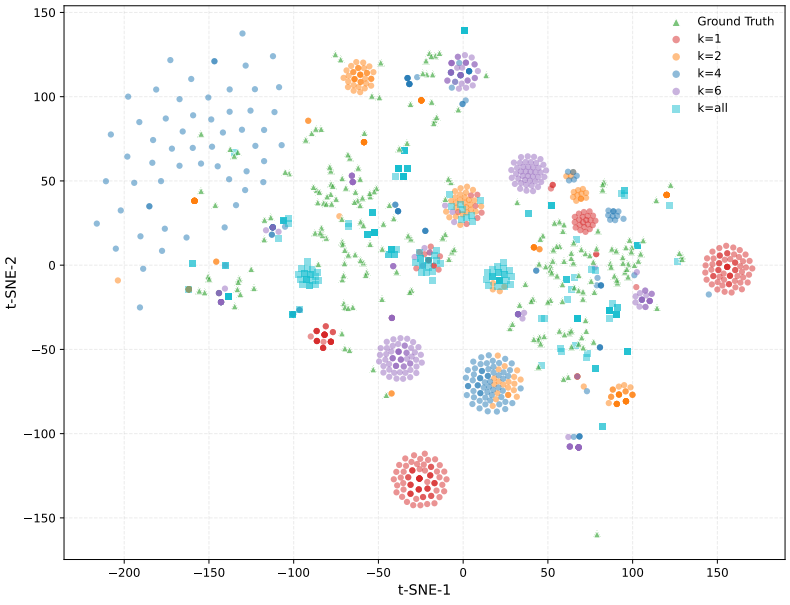


Figure 8: t-SNE visualization of case-level representations under different input-view budgets $K \in \{1, 2, 4, 6, \text{All}\}$.