VESP: VISUAL EVIDENCE AND SEMANTIC CONSISTENCY PRESERVATION IN AUTOREGRESSIVE MRI REPORT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Automated radiology report generation from 3D MRI remains challenging because visual evidence must remain associated with the appropriate clinical findings and influential throughout autoregressive generation. Existing approaches typically introduce visual information during encoding, but may suffer from insufficient field-level evidence organization, semantic inconsistency between visual and textual findings, and diminishing visual influence as generation progresses.

We propose VESP, a visual evidence preservation framework for structured MRI report generation that maintains image-grounded evidence from visual encoding to final token prediction. First, Schema-guided Visual Evidence Linking (SVEL) organizes shared MRI representations into field-specific visual memories and propagates information among related clinical fields using schema-derived relations. Second, Visual-Text Evidence Alignment (VTEA) aligns image-side and report-side predictions within a shared structured finding space, encouraging generated findings to remain consistent with visual evidence. Third, Evidence-Retained Adaptive Decoding (ERAD) selectively reuses learned field-level visual predictions at schema-defined candidate decisions during inference, preserving visual influence without additional training or learnable parameters.

Experiments on a private shoulder MRI dataset and the public KneeCoT knee MRI dataset show consistent improvements, achieving 75.65 micro-F1 on shoulder MRI and 96.69 micro-F1 on knee MRI. These results demonstrate that preserving evidence identity, semantics, and influence throughout autoregressive generation improves the accuracy and image consistency of structured MRI reports.

1 INTRODUCTION

Radiology report generation aims to translate medical images into clinically meaningful descriptions. Beyond producing fluent text, a reliable model must preserve the correspondence between visual evidence and the clinical findings expressed in the report. This requirement is particularly challenging for volumetric MRI, where an examination may contain hundreds of slices while clinically relevant abnormalities often occupy only a small fraction of the volume. Therefore, successful report generation requires not only extracting informative visual representations, but also maintaining visual evidence throughout the complete image-to-report generation process.

Recent advances have extended medical report generation from 2D radiography toward volumetric vision-language modeling. CT2Rep introduced an end-to-end framework for generating reports from 3D CT by combining volumetric visual encoding with autoregressive decoding (Hamamci et al., 2024). Med3DVLM further explored efficient 3D medical vision-language modeling by integrating volumetric representation learning with large language models (Xin et al., 2025). More recently, MRI2Rep proposed structured autoregressive reporting for 3D MRI by converting free-form reports into structured diagnostic sequences (Li et al., 2026). Despite these advances, existing approaches mainly focus on improving visual representation learning or language generation, while how to preserve image-grounded evidence from visual encoding to final token prediction remains insufficiently explored.

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.

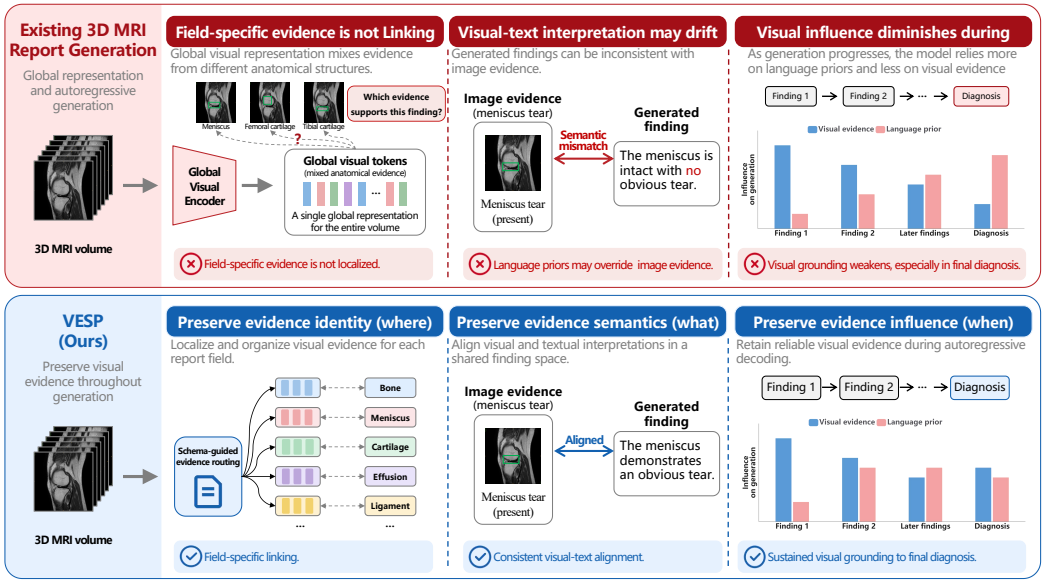


Figure 1: Motivation of visual evidence preservation for structured MRI report generation. Visual evidence can be progressively lost during image-to-report generation: (1) evidence identity is not explicitly organized across clinical fields, (2) visual and textual interpretations may become inconsistent, and (3) visual influence diminishes during long autoregressive decoding. VESP addresses these failures by preserving evidence identity, semantics, and influence throughout the generation pathway.

We argue that visual evidence is progressively lost during image-to-report generation. This evidence degradation occurs at three interconnected stages: preserving which evidence belongs to each clinical field, preserving the semantic consistency between visual evidence and generated findings, and preserving the influence of visual evidence during long autoregressive decoding.

First, visual evidence identity is difficult to preserve across clinical fields. Structured musculoskeletal MRI reports contain multiple anatomical sections and findings, where different fields depend on different visual information. However, existing approaches typically rely on global visual representations, which encode examination-level characteristics but do not explicitly maintain field-specific evidence identity. Consequently, the language model receives visual information without a clear correspondence between clinical fields and their supporting evidence. Although recent studies have explored regional supervision, anatomy-aware generation, and structured clinical grounding (Chen et al., 2024; Di Piazza et al., 2025; Gao et al., 2026), these approaches do not explicitly organize heterogeneous volumetric information into field-specific visual representations for structured report generation.

Second, visual evidence semantics may drift during report generation. Autoregressive vision-language models are optimized to produce coherent sequences, but the generation objective alone does not explicitly enforce faithfulness to visual observations. Similar issues have been observed in large vision-language models, where models may generate plausible but unsupported descriptions due to over-reliance on language priors (Leng et al., 2024). As a result, models may rely on linguistic patterns, reporting templates, or frequent clinical descriptions and generate plausible findings that deviate from the underlying MRI observations. Reliable report generation therefore requires an explicit mechanism to maintain consistency between visual interpretations and textual findings.

Third, visual evidence influence may diminish during autoregressive decoding. Even when visual information is incorporated during encoding or multimodal alignment, later token predictions may increasingly depend on previously generated text and language priors. Recent studies on vision-language hallucination have shown that decoding strategies can improve generation faithfulness by restoring or strengthening visual influence (Leng et al., 2024; Huang et al., 2024).

To address these challenges, we propose **VESP**, a visual evidence preservation framework for structured musculoskeletal MRI report generation. Rather than treating visual grounding as a single-stage representation learning problem, VESP establishes a continuous evidence preservation pathway that maintains image–report correspondence from visual encoding to final token prediction.

Specifically, VESP preserves three complementary properties of visual evidence. **Schema-guided Visual Evidence Localization (SVEL)** preserves evidence identity by treating the clinical reporting schema as an evidence routing interface. Field-level queries retrieve dedicated visual memories for individual reporting fields, while relation-aware propagation models dependencies among related clinical fields. **Visual-Text Evidence Alignment (VTEA)** preserves evidence semantics by jointly optimizing image-side and report-side predictions within the same clinical finding space, encouraging generated findings to remain compatible with visual evidence. Finally, **Evidence-Retained Adaptive Decoding (ERAD)** preserves evidence influence during inference by selectively reusing learned field-level visual predictions at schema-defined token decisions, without additional training or learnable parameters.

We evaluate VESP on two musculoskeletal MRI domains, including multi-sequence shoulder MRI and single-sequence knee MRI. Experimental results demonstrate consistent improvements over existing volumetric report-generation approaches. Further analysis shows that the improvement becomes more pronounced in later report sections, where maintaining visual evidence is particularly challenging for autoregressive generation.

The contributions of this work are summarized as follows. We identify visual evidence preservation as a fundamental challenge in structured medical report generation and characterize three forms of evidence degradation during the image-to-report process: identity loss, semantic drift, and influence decay. We propose VESP, a continuous evidence preservation framework that maintains visual evidence through schema-guided routing, visual-text alignment, and evidence-retained decoding. Extensive experiments on shoulder and knee MRI demonstrate the effectiveness of preserving visual evidence throughout structured medical report generation.

2 Related Work

Volumetric medical vision-language generation. Medical report generation has progressed from 2D radiography toward volumetric medical imaging. Early methods mainly focused on image-text alignment and language generation for chest X-rays (Wang et al., 2018; Chen et al., 2020; 2021).

Recent volumetric models enable direct report generation from 3D images. CT2Rep combines volumetric visual encoding with autoregressive decoding for 3D CT reporting (Hamamci et al., 2024). Med3DVLM integrates volumetric representation learning with large language models (Xin et al., 2025), while MRI2Rep formulates 3D MRI reporting as structured autoregressive generation (Li et al., 2026). Despite these advances, existing approaches mainly improve visual representations or generation capability. A global visual representation does not explicitly identify which evidence supports individual clinical findings, motivating visual evidence organization according to clinical reporting structures.

Structured clinical grounding and evidence alignment. Structured and anatomy-aware representations have been explored to improve image-report correspondence. Structured Radiology Report Generation reformulates narrative reports into standardized clinical fields (Delbrouck et al., 2025). RadGraph extracts clinical entities and relations for structured report analysis (Jain et al., 2021), while CheXbert extracts structured findings for medical report supervision and evaluation (Smit et al., 2020).

Anatomy-aware generation further introduces regional or clinical guidance. X-RGen uses cross-region supervision (Chen et al., 2024), anatomy-guided generation incorporates regional anatomical prompts (Gao et al., 2026), CT-GRAPH models hierarchical anatomical relations in volumetric imaging (Kalisch et al., 2025), and LLaVA-TA decomposes reports into clinically meaningful topics (Cheng & Subramanian, 2026). These methods demonstrate the value of explicit clinical structure, but image-derived and text-derived findings are generally not explicitly constrained to agree. In contrast, VESP directly aligns their predictions within a shared structured finding space.

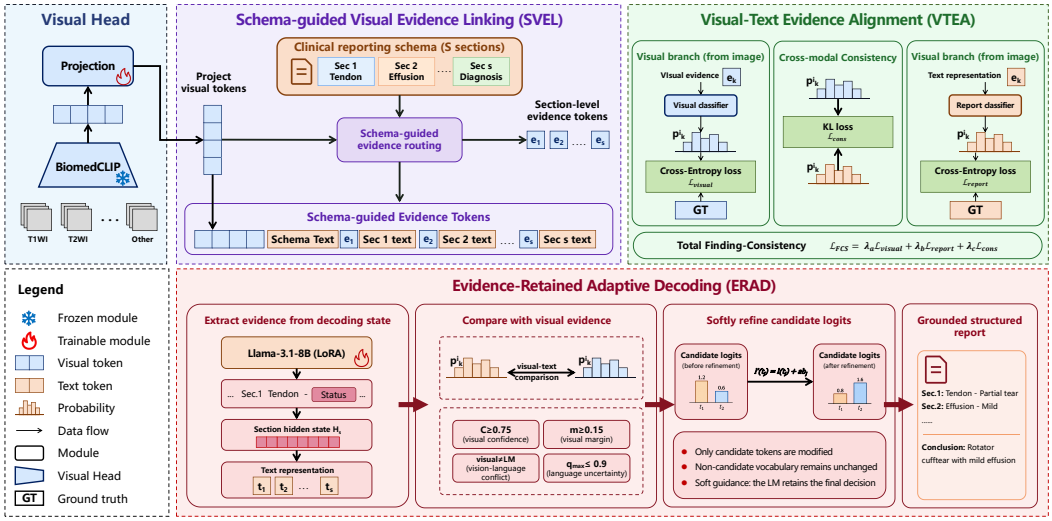


Figure 2: Overview of VESP. The framework preserves visual evidence across three stages: SVEL organizes shared MRI representations into field-specific evidence, VTEA aligns visual and textual interpretations in a shared finding space, and ERAD selectively reuses visual predictions during autoregressive decoding.

Preserving visual evidence during autoregressive decoding. Multimodal language models connect visual encoders and language models through cross-modal interfaces such as cross-attention and query-based alignment (Alayrac et al., 2022; Li et al., 2023a). Learned visual queries have also been widely used to retrieve task-relevant representations (Carion et al., 2020).

However, visual grounding can weaken during autoregressive generation as models increasingly rely on language priors. POPE highlights visually unsupported generation in large vision-language models (Li et al., 2023b). Inference-time methods such as Visual Contrastive Decoding (Leng et al., 2024) and OPERA (Huang et al., 2024) therefore modify decoding to strengthen visual reliance.

These methods mainly operate on global visual representations and general vision-language tasks. In contrast, VESP learns field-specific visual evidence during structured training and reuses it at schema-defined decoding decisions, connecting evidence organization, cross-modal alignment, and decoding-time preservation within a unified pathway.

3 Method

3.1 Framework Overview

VESP preserves visual evidence throughout structured MRI report generation by maintaining three complementary properties: *identity*, *semantics*, and *influence*. Given an MRI examination, each sequence is encoded by a frozen BiomedCLIP encoder (Zhang et al., 2023) and projected into the embedding space of Llama-3.1-8B (Grattafiori et al., 2024), producing shared visual tokens V.

To preserve evidence identity, **Schema-guided Visual Evidence Linking (SVEL)** uses the clinical reporting schema to organize shared MRI representations into field-specific visual memories. To preserve evidence semantics, **Visual-Text Evidence Alignment (VTEA)** constrains visual-side and report-side interpretations within the same structured finding space. Finally, to preserve evidence influence, **Evidence-Retained Adaptive Decoding (ERAD)** reuses learned visual finding predictions at schema-defined candidate branching points during inference.

3.2 Schema-guided Visual Evidence Linking (SVEL)

A global visual representation contains information from the complete MRI examination but does not explicitly provide each reporting field with a dedicated evidence representation. SVEL therefore

treats the reporting schema as an intermediate evidence-organization space rather than attempting precise anatomical localization.

For each reporting field f, a learnable query retrieves a field-conditioned visual memory from the shared visual tokens, following standard query-based visual retrieval (Carion et al., 2020; Li et al., 2023a):

$$m_f = \mathrm{CrossAttn}(q_f, V, V), \tag{1}$$

where q_f is the query associated with field f and m_f is its dedicated visual memory.

To model dependencies among report fields, we construct a relation graph directly from the clinician-defined schema. The graph contains self-relations, relations among fields within the same report section, and predefined Findings-to-Conclusion relations. Relation-aware attention propagates information among connected field memories:

$$\tilde{m}_f = m_f + \mathrm{RelAttn}\big(\{m_g : g \in \mathcal{N}(f)\}\big). \tag{2}$$

The enhanced field memories are subsequently aggregated within each report section and inserted together with schema descriptions as structured visual evidence for the language model. Thus, SVEL establishes field-to-evidence correspondence without requiring explicit lesion localization.

3.3 Visual-Text Evidence Alignment (VTEA)

Field-specific evidence does not by itself guarantee that generated findings remain semantically consistent with the image. VTEA therefore supervises visual and textual interpretations within the same predefined finding space.

For field f, the visual branch predicts from $\tilde{m}_f$, while the text branch predicts from the corresponding language-model representation H_f:

$$p_f^v = \mathrm{softmax}(h_f^v(\tilde{m}_f)), \qquad p_f^t = \mathrm{softmax}(h_f^t(H_f)). \tag{3}$$

The two branches are trained with structured finding supervision and an agreement constraint:

$$\mathcal{L}_{\mathrm{VTEA}} = \lambda_v \mathcal{L}_v + \lambda_t \mathcal{L}_t + \lambda_c D_{\mathrm{KL}}(p_f^v \| p_f^t). \tag{4}$$

This shared finding space encourages textual findings to remain compatible with image-derived evidence rather than relying solely on language priors.

3.4 Evidence-Retained Adaptive Decoding (ERAD)

Even after evidence organization and alignment, visual influence can weaken during long autoregressive generation. ERAD reuses the learned visual finding predictions only at structured clinical decision points during inference.

When the generated prefix reaches a schema-defined field, the valid structured options are mapped to their possible next-token branches. If multiple complete options share the same next token, they are grouped into one branch. For candidate branch j, the corresponding visual score is

$$s_j = \frac{1}{|G_j|} \sum_{i \in G_j} v_i, \tag{5}$$

where G_j denotes the structured options associated with candidate token t_j, and v_i is the visual prediction logit for option i.

The candidate scores are centered and clipped to produce a bounded visual bias, which is added only to candidate logits:

$$l'(t_j) = l(t_j) + \alpha \ \mathrm{clip}\big(s_j - \mathrm{mean}(s), -2, 2\big), \qquad \alpha = 0.1. \tag{6}$$

All non-candidate vocabulary logits remain unchanged. Therefore, ERAD is not a global fusion of an auxiliary classifier with the language model. It performs lightweight, candidate-specific visual reranking only at schema-defined decision boundaries and requires no additional training or learnable parameters.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate VESP on two musculoskeletal MRI domains: a private shoulder MRI dataset and the publicly available KneeCoT knee MRI dataset (Niu et al., 2026). The two domains differ in anatomical structures, imaging configurations, and reporting schemas, providing complementary settings for evaluating visual evidence preservation.

For the shoulder domain, the training, validation, and held-out test sets contain 5,819, 737, and 747 examinations, respectively. Each examination may contain multiple MRI sequences or weightings. The shoulder reporting schema contains 60 structured fields organized into nine anatomical sections.

For the knee domain, we use KneeCoT (Niu et al., 2026). The training, validation, and held-out test sets contain 10,536, 1,316, and 1,317 examinations, respectively. Unlike the shoulder domain, each examination contains a single MRI sequence. The knee reporting schema contains 67 structured fields organized into 14 anatomical sections.

Each domain uses a clinician-defined reporting schema specifying anatomical sections, field definitions, and candidate finding options. The same schema serves as the interface for field-specific evidence organization, visual-text alignment, and structured evaluation. During generation, only the schema definition is available to the model; patient-specific structured labels are used only for supervision and evaluation.

The original free-text reports are converted into structured field-level annotations according to the predefined schemas. These annotations supervise visual finding prediction and visual-text evidence alignment and are also used for structured evaluation.

The visual encoder is initialized with BiomedCLIP and remains frozen during training. Visual representations are projected into the embedding space of Llama-3.1-8B-Instruct, which is adapted using LoRA. All model configurations, hyperparameters, and checkpoints are selected using the validation set based on micro-F1. The held-out test set is used only for final evaluation, including the reported ablation analyses.

We evaluate generated reports at both structured-field and text levels. The primary metrics are schema-based micro precision, recall, and F1 over all structured report fields. We additionally report BLEU-1, BLEU-4, ROUGE-L, and METEOR to measure textual similarity between generated and reference reports. Invalid field predictions outside the predefined schema option space are also monitored.

For the section-level analysis in Sec. 4.4, we compare the micro-F1 of VESP with the base model separately across report sections. This analysis examines whether the benefit of visual evidence preservation changes from early Findings sections to later Diagnosis sections.

4.2 COMPARISON WITH EXISTING METHODS

We compare VESP with representative volumetric medical vision-language and report-generation approaches, including CT2Rep (Hamamci et al., 2024), Med3DVLM (Xin et al., 2025), BTB3D (Hamamci et al., 2025), MRI2Rep (Li et al., 2026), M3D-LaMed (Bai et al., 2024), and MedGemma (Sellergren et al., 2026).

Tables 1 and 2 summarize the results on the private shoulder MRI dataset and KneeCoT, respectively.

On shoulder MRI, VESP achieves 75.65 micro-F1, outperforming MRI2Rep, the strongest baseline on this domain, by 2.13 percentage points. On KneeCoT, VESP obtains 96.69 micro-F1, improving

Table 1: Comparison with representative volumetric medical vision-language and report-generation baselines on the private shoulder MRI test set. P, R, and F1 denote schema-based micro precision, recall, and F1, respectively.

Method	P	R	F1	BLEU-1	BLEU-4	ROUGE-L	METEOR	F1 95% CI
M3D-LaMed $_{arXiv'24}$	**82.67**	49.83	62.18	0.5606	0.5375	0.7635	0.6558	61.02–63.21
CT2Rep $_{MICCAI'24}$	74.53	65.11	69.50	0.3412	0.3174	0.4853	0.7543	68.31–70.59
BTB3D $_{NeurIPS'25}$	71.37	61.73	66.20	0.8457	0.7924	0.9084	0.8609	64.95–67.31
Med3DVLM $_{\text{JBHI}'26}$	58.00	56.46	57.22	0.8346	0.7288	0.8100	0.8092	56.17–58.10
MRI2Rep $_{MICCAI'26}$	77.63	69.83	73.52	0.8882	0.8421	0.9227	0.8939	72.32–74.73
MedGemma $_{arXiv'26}$	75.42	66.66	70.77	0.6586	0.6111	0.7389	0.8684	69.80–72.16
VESP	80.19	**71.59**	**75.65**	**0.8984**	**0.8545**	**0.9287**	**0.9027**	**74.61–76.64**

Table 2: Comparison with representative volumetric medical vision-language and report-generation baselines on the KneeCoT knee MRI test set. P, R, and F1 denote schema-based micro precision, recall, and F1, respectively.

Method	P	R	F1	BLEU-1	BLEU-4	ROUGE-L	METEOR	F1 95% CI
M3D-LaMed $_{arXiv'24}$	95.36	58.10	72.20	0.4747	0.4724	0.7241	0.5975	71.87–72.56
CT2Rep $_{MICCAI'24}$	87.52	92.97	90.17	0.1749	0.1734	0.3110	0.6847	89.69–90.67
BTB3D $_{NeurIPS'25}$	94.44	94.39	94.41	0.9931	0.9867	0.9918	0.9927	93.98–94.85
Med3DVLM $_{JBHI'26}$	93.75	94.78	94.26	0.4052	0.4034	0.5741	0.8706	93.82–94.67
MRI2Rep $_{MICCAI'26}$	91.94	91.90	91.92	0.9914	0.9826	0.9883	0.9906	91.47–92.38
MedGemma $_{arXiv'26}$	86.49	86.41	86.45	0.7370	0.7225	0.8388	0.9474	86.10–86.78
VESP	**96.71**	**96.65**	**96.69**	**0.9942**	**0.9893**	**0.9936**	**0.9935**	**96.41–96.95**

over the strongest baseline BTB3D (94.41) by 2.28 percentage points. VESP also achieves strong text-generation performance across both domains.

Notably, the relative ranking of individual baselines changes substantially between the two datasets. For example, MRI2Rep achieves 73.52 F1 on shoulder MRI compared with 35.58 for Med3DVLM, whereas on KneeCoT Med3DVLM reaches 94.26 F1 compared with 91.92 for MRI2Rep. The shoulder degradation of Med3DVLM is primarily associated with low recall (24.48), rather than a uniform reduction in both precision and recall.

We attribute this ranking variation cautiously to the interaction between model design and domain-specific task structure rather than to a single difference in dataset difficulty. The private shoulder dataset contains multi-sequence MRI and a different structured reporting schema, whereas KneeCoT contains single-sequence knee MRI with a distinct field organization. Models designed for general volumetric vision-language modeling and models explicitly designed for structured autoregressive reporting may therefore respond differently to these settings. Importantly, VESP remains consistently strong despite this change in baseline behavior, suggesting that the benefit of evidence preservation is not tied to a particular imaging configuration or baseline family.

4.3 Ablation of the Visual Evidence Preservation Pathway

We conduct progressive ablation experiments on the 747-case held-out shoulder test set to evaluate the contribution of the three stages of the proposed visual evidence preservation pathway: Schema-guided Visual Evidence Linking (SVEL), Visual-Text Evidence Alignment (VTEA), and Evidence-Retained Adaptive Decoding (ERAD).

The base model retains global visual conditioning but removes all three evidence-preservation mechanisms. SVEL introduces field-specific visual memories and schema-derived relation propagation. VTEA aligns visual and textual predictions within a shared structured finding space. ERAD reuses learned visual predictions during inference to refine schema-defined candidate decisions without additional training.

The progressive ablation results in Table 3 support the three-stage evidence preservation design. Starting from global visual conditioning, SVEL improves micro-F1 from 73.36 to 73.48, indicating

Table 3: Ablation of the visual evidence preservation pathway on the 747-case private shoulder MRI test set.

SVEL	VTEA	ERAD	P	R	F1
×	×	×	78.47	68.88	73.36
✓	×	×	78.49	69.08	73.48
✓	✓	×	78.96	70.01	74.22
✓	✓	✓	80.18	71.59	75.64

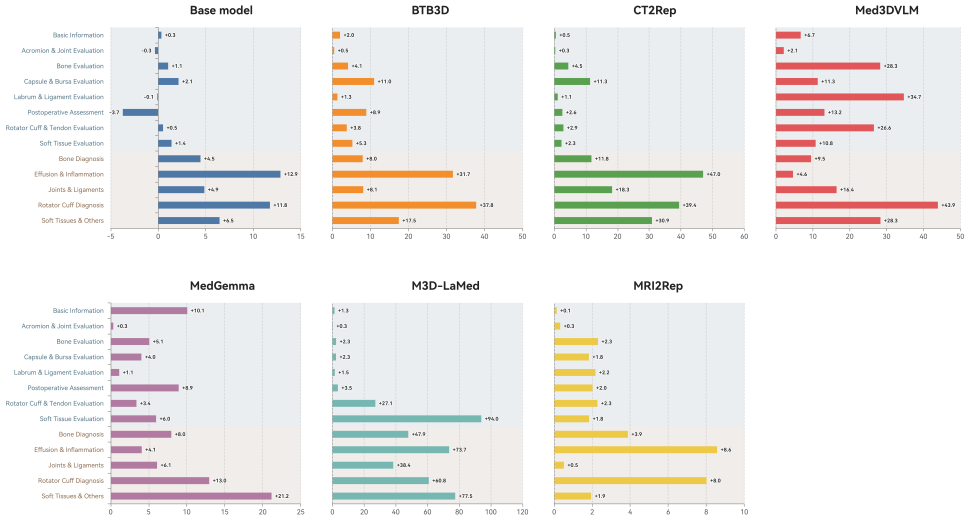


Figure 3: Section-level micro-F1 improvements of VESP over each baseline on the private shoulder MRI test set. The upper-left panel shows the improvement over the ablation base model; the remaining panels show improvements over the external baselines.

that organizing shared visual representations into field-specific evidence provides additional information beyond global visual tokens.

Adding VTEA further increases micro-F1 to 74.22, suggesting that explicitly constraining visual and textual interpretations provides additional benefit after field-specific evidence organization. The complete model with ERAD achieves 75.64 micro-F1. Since ERAD operates only during inference without modifying trained model parameters, the final gain indicates that previously learned visual predictions can provide additional value when reused at structured decoding decisions.

4.4 Evidence Preservation Across Report Sections

The overall ablation results do not reveal where the improvement occurs within the structured report. We therefore compare section-level micro-F1 between VESP and each baseline on the shoulder test set, as shown in Fig. 3.

The improvement is strongly non-uniform across report sections. Relative to the base model, most Findings sections show relatively modest differences, including Basic Information (+0.3 pp), Bone Evaluation (+1.1 pp), Capsule & Bursa Evaluation (+2.1 pp), and Soft Tissue Evaluation (+1.4 pp), while Acromion & Joint Evaluation (-0.3 pp), Labrum & Ligament Evaluation (-0.1 pp), and Postoperative Assessment (-3.7 pp) show small negative differences. In contrast, substantially larger gains occur in the later Diagnosis sections: Effusion & Inflammation (+12.9 pp), Rotator Cuff Diagnosis (+11.8 pp), Soft Tissues & Others (+6.5 pp), Joints & Ligaments (+4.9 pp), and Bone Diagnosis (+4.5 pp).

The same pattern holds, with even larger magnitude, against the external baselines. Over the strongest baseline, MRI2Rep, gains remain modest in most Findings sections (below +2.5 pp) but reach +8.0 pp in Rotator Cuff Diagnosis and +8.6 pp in Effusion & Inflammation; against weaker baselines the Diagnosis-section gains are substantially larger, exceeding +40 pp for Med3DVLM and +70 pp for M3D-LaMed, indicating that the later-stage benefit is not specific to a particular baseline family.

This pattern is consistent with the motivation of visual evidence preservation. Early Findings fields are generated with relatively short autoregressive context, whereas Diagnosis sections occur later and summarize evidence distributed across preceding findings, making these decisions more exposed to accumulated linguistic context and potential loss of visual influence. The Diagnosis-section gains therefore provide additional evidence that preserving visual information becomes particularly important during later stages of autoregressive report generation.

5 Discussion

The experiments reveal several insights beyond the overall performance gains of VESP. First, the benefit of visual evidence preservation depends on where a decision occurs in the report. Gains are modest across most Findings sections but become substantially larger in later Diagnosis sections. Early predictions are made with limited textual history, whereas later diagnostic decisions must remain grounded in image evidence after considerably more linguistic context has accumulated. Visual evidence preservation therefore becomes increasingly important as generation progresses. Second, field-specific visual organization alone is insufficient. SVEL provides only modest gains over global visual conditioning, while larger improvements appear after introducing VTEA and ERAD. This shows that separating visual evidence by clinical field is useful, but the evidence must also remain semantically aligned with generated findings and influential during subsequent structured decisions. Evidence preservation is therefore better viewed as a generation-wide process rather than solely a representation-learning problem. Third, existing methods show markedly different behavior across shoulder and knee MRI, whereas VESP remains consistently strong in both settings. This variation suggests that volumetric vision-language architectures may be sensitive to domain-specific factors such as imaging configuration and reporting structure. Organizing and reusing visual evidence through the reporting schema provides a more consistent mechanism across heterogeneous structured-reporting settings. Several limitations remain. The section-level analysis reflects the combined effect of the full preservation pathway and does not isolate ERAD's contribution to individual late-stage sections. Evaluation is limited to private shoulder MRI and public KneeCoT knee MRI, motivating broader validation across anatomical regions, institutions, and imaging protocols. In addition, VESP relies on predefined reporting schemas that require clinical knowledge to construct. Finally, the current evaluation mainly uses automatic structured and textual metrics; radiologist assessment, clinically weighted error analysis, and uncertainty-aware evaluation remain important directions for future work.

6 Conclusion

We present VESP, a framework for preserving visual evidence throughout autoregressive medical report generation. Rather than relying on visual conditioning only at the input stage, VESP maintains field-specific evidence through schema-guided linking, visual-text alignment, and evidence-retained decoding.

Experiments on private shoulder MRI and public KneeCoT knee MRI demonstrate consistent improvements across heterogeneous reporting settings. More importantly, the gains become substantially larger in later Diagnosis sections, indicating that maintaining image-grounded evidence becomes increasingly important as autoregressive context grows.

These results suggest that reliable medical report generation depends not only on stronger visual representations or language models, but also on preserving visual evidence throughout the generation trajectory.

AI USE STATEMENT

Generative AI tools were used in this work only for dataset construction and language polishing. During dataset construction, large language models were used to convert free-text radiology reports into field-level annotations under the predefined clinical reporting schemas. Specifically, three language models independently mapped each report to the candidate option space of each field, and agreement across models was used to obtain structured supervision; unresolved cases were manually reviewed according to the data-construction protocol described in Sec. 4.1. These AI-assisted annotations were used only for structured supervision and evaluation.

For language polishing, generative AI was used only to improve the language, grammar, and readability of the manuscript. All AI-assisted outputs were reviewed by the authors. Generative AI was not used to develop research hypotheses, design methodology, conduct or interpret experimental analyses, perform literature search, design figures, or generate MRI images or reported numerical experimental results. The authors take full responsibility for the final content of this work, including text, annotations, claims, and artifacts.

REFERENCES

Jean-Baptiste Alayrac et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022.

Fan Bai, Yuxin Du, Tiejun Huang, Max Q.-H. Meng, and Bo Zhao. M3D: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020*, pp. 213–229. Springer, 2020. doi: 10.1007/978-3-030-58452-8_13.

Qi Chen, Yutong Xie, Biao Wu, Xiaomin Chen, James Ang, Minh-Son To, Xiaojun Chang, and Qi Wu. Act like a radiologist: Radiology report generation across anatomical regions. In *Asian Conference on Computer Vision (ACCV)*, 2024.

Zhihong Chen, Yan Song, Yu Chang, and Xia Wan. Generating radiology reports via memory-driven transformer. In *EMNLP*, 2020.

Zhihong Chen, Yan Shen, Yan Song, and Xia Wan. Cross-modal memory networks for radiology report generation. In *ACL-IJCNLP*, 2021.

Sheng Cheng and Devika Subramanian. Rethinking radiology report generation: From narrative flow to topic-guided findings. In *International Conference on Learning Representations (ICLR)*, 2026.

Jean-Benoit Delbrouck, Justin Xu, Johannes Moll, Alois Thomas, Zhihong Chen, Sophie Ostmeier, Asfandyar Azhar, Kelvin Zhenghao Li, Andrew Johnston, Christian Bluethgen, Eduardo Reis, Mohamed Muneer, Maya Varma, and Curtis Langlotz. Automated structured radiology report generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26813–26829, 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)*, 2025. doi: 10.1109/ISBI60581.2025.10981073.

Yijian Gao, Dominic Marshall, Xiaodan Xing, Junzhi Ning, Congren Dai, Giorgos Papanastasiou, Guang Yang, and Matthieu Komorowski. Anatomy-guided radiology report generation with pathology-aware regional prompts. *IEEE Open Journal of Engineering in Medicine and Biology*, 7:165–171, 2026. doi: 10.1109/OJEMB.2026.3687122.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. CT2Rep: Automated radiology report generation for 3d medical imaging. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pp. 476–486. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72390-2_45.

Ibrahim Ethem Hamamci, Sezgin Er, Suprosanna Shit, Hadrien Reynaud, Dong Yang, Pengfei Guo, Marc Edgar, Daguang Xu, Bernhard Kainz, and Bjoern Menze. Better tokens for better 3d: Advancing vision-language modeling in 3d medical imaging. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4507.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024. doi: 10.1109/CVPR52733.2024.01274.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, et al. Radgraph: Extracting clinical entities and relations from radiology reports. In *NeurIPS*, 2021.

Hamza Kalisch, Fabian Hörst, Jens Kleesiek, Ken Herrmann, and Constantin Seibold. CT-GRAPH: Hierarchical graph attention network for anatomy-guided ct report generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 6834–6843, 2025. doi: 10.1109/ICCVW69036.2025.00706.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024. doi: 10.1109/CVPR52733.2024.01316.

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 *ICML*, 2023a.

Xinran Li, Junlin Yang, Annabella Shewarega, Zongwei Zhou, Julius Chapiro, James S. Duncan, and Lawrence H. Staib. MRI2Rep: Autoregressive structured report generation for 3d liver mri. *arXiv preprint arXiv:2606.25279*, 2026.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating and mitigating object hallucination in large vision-language models. In *EMNLP*, 2023b.

Ben Niu, Yihui Li, Kai Guo, Yuli Wang, Jiaqi Xie, and Along He. Kneecot: A large-scale chain-of-thought benchmark for knee mri diagnostic vision-language models. In *ACM Conference Proceedings*, 2026.

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, et al. MedGemma 1.5 technical report. *arXiv preprint arXiv:2604.05081*, 2026.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1500–1519. Association for Computational Linguistics, 2020.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *CVPR*, 2018.

Yu Xin, Gorkem Can Ates, Kuang Gong, and Wei Shao. Med3DVLM: An efficient vision-language model for 3d medical image analysis. *arXiv preprint arXiv:2503.20047*, 2025.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, et al. BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.