Glance Before You Tell: Anomaly-Guided 3D Radiology Report Generation with Heat-Conduction Slice Encoders

Lingyu Zhou
College of Computer Science
Sichuan University
Chengdu, Sichuan, China
scu_zly@stu.scu.edu.cn

Dingwen Pi
College of Computer Science
Sichuan University
Chengdu, Sichuan, China
pdwen@stu.scu.edu.cn

Zhang Yi
College of Computer Science
Sichuan University
Chengdu, Sichuan, China
zhangyi@scu.edu.cn

Jun Wu
College of Computer Science
Sichuan University
Chengdu, Sichuan, China
junwu@stu.scu.edu.cn

Xiuyuan Xu*
School of Artificial Intelligence
Sichuan University
Chengdu, Sichuan, China
xuxiuyuan@scu.edu.cn

Abstract

Generating accurate radiology reports from 3D medical volumes remains prone to hallucination. Existing grounding strategies often depend on external visual foundation models or on fixed medical phrase queues. The former passes segmenter-derived evidence to generation, where segmentation errors are hard to suppress. The latter has limited coverage of long-tail and compositional abnormalities. We therefore propose HEATRAD, a *glance-before-tell* framework that first routes suspicious slice evidence into an organ-ordered visual prefix, then generates reports from this localized evidence without explicit grounding supervision. HEATRAD uses a slice-centric signed heat-conduction encoder that performs efficient spectral mixing while separating smooth anatomy from lesion-like residues. HEATRAD then scores suspicious slices, binds their evidence to coarse organ contexts, and assembles an organ-ordered visual prefix, guiding the decoder toward localized abnormalities instead of redundant volumetric context. Experiments on three 3D radiology benchmarks show consistent gains in clinical fidelity and hallucination-related metrics, supporting anomaly-first evidence selection as an annotation-efficient paradigm for 3D report generation.

Code: https://github.com/SCU-zly/HeatRAD

1 Introduction

3D computed tomography (CT) volumes are central to medical diagnosis, yet the growth of imaging workload continues to outpace radiological capacity (Rajpurkar et al., 2022). Recent multimodal large language models (MLLMs) (Liu et al., 2023; Alayrac et al., 2022; Li et al., 2023a) have therefore stimulated rapid progress in 3D radiology report generation (3D RRG), where systems draft reports directly from volumetric scans (Hamamci et al., 2024a; Bai et al., 2024a; Wu et al., 2025; Hamamci et al., 2024b; Chen et al., 2025a). A persistent obstacle to clinical use, however, is *hallucination*: the model may describe findings unsupported by the input volume, such as nonexistent nodules, effusions,

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

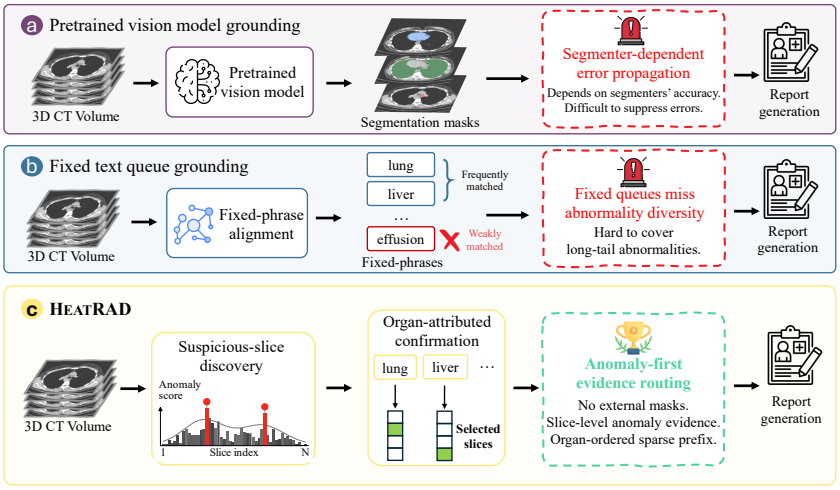

Figure 1: HeatRAD avoids external segmenters and fixed queues by discovering suspicious slices, attributing organ-level evidence, and forming a sparse prefix for report generation.

or anatomical variants, while maintaining fluent and clinically plausible language (Li et al., 2023b; Bai et al., 2024b; Huang et al., 2024; Leng et al., 2024). In radiology, such unsupported findings are not merely linguistic errors; they alter the evidential basis of diagnosis. A natural way to constrain hallucination is thus to make report generation conditional on explicit visual evidence, so that each generated abnormality can be traced back to the volume region that supports it.

Existing grounding strategies in 3D RRG mainly instantiate this idea in two ways. One line injects evidence from external pretrained vision models, such as large-scale organ segmenters or SAM-style backbones (Wasserthal et al., 2023; Zhao et al., 2023; Kirillov et al., 2023; Ma et al., 2024), and conditions the report decoder on their predicted masks or regions (Tanida et al., 2023; Bannur et al., 2024). This design provides spatial structure, however, grounding generation on strong pseudo labels makes the model highly dependent on segmenters' accuracy, allowing segmentation errors or missed regions to propagate into the language model and making them difficult to suppress once verbalized as apparently grounded findings. Another line aligns visual regions to a curated queue of medical entity phrases, typically fixed organ names and lesion descriptors (Zhang et al., 2024; Hyland et al., 2023; Müller et al., 2024). Such phrase-level grounding is appealing in form, but radiological abnormalities are highly long-tailed: common templates are repeatedly matched, whereas rare or institution-specific descriptions are underrepresented. As a result, fixed queues tend to over-trigger frequent findings and miss rare ones, increasing both false positives and false negatives. The resulting evidence–text alignment is therefore fragile; in difficult cases, grounding can amplify hallucination rather than suppress it.

Radiologists often follow an anomaly-first workflow: they inspect the volume for suspicious slices, interpret each suspicious observation within its anatomical context, and then verbalize the supported findings. This motivates the principle of *glance before tell*: report generation should be conditioned on a prior pass that identifies suspicious visual evidence before language decoding. We instantiate this principle with HeatRAD, an anomaly-first framework that performs annotation-free evidence routing over CT volumes. Figure 1 contrasts this design with prior grounding pipelines. Concretely, HeatRAD contains three end-to-end coupled components. First, a *Heat-Conduction Slice Encoder* performs efficient frequency-conditioned token mixing within each slice through a learnable signed spectral gate, which can attenuate low-frequency anatomical baselines while amplifying lesion-like high-frequency residues; axial continuity is injected only at the deepest level to preserve the per-slice addressability needed for evidence routing. Second, a suspicious-slice discovery module compares each slice representation with volume-conditioned normal/abnormal anchors (Xu et al., 2025), yielding a continuous anomaly probability that gates organ attribution. Third, an *Organ-Attributed Confirmation* module uses frozen biomedical text embeddings (Shui et al., 2025) as organ anchors, binding suspicious evidence to stable anatomical contexts without external masks. An organ-wise sparsemax sampler (Martins & Astudillo, 2016) with a learnable

abstain anchor then selects sparse, organ-consistent evidence and assembles it into a structured visual prefix, from which the language model generates the final report. Our contributions are threefold:

- **An anomaly-first formulation for 3D RRG.** We decompose report generation into slice-level anomaly scoring, organ-attributed evidence grouping, and evidence-conditioned decoding, trained end-to-end on volume–report pairs without task-specific spatial supervision.
- **A signed slice-centric heat-conduction encoder.** We introduce signed in-plane HCO to suppress smooth anatomy while amplifying lesion-like residues, and add lightweight inter-slice HCO to inject axial context while preserving slice addresses.
- **Text-anchored organ attribution with adaptive sparse selection.** We bind suspicious slices to frozen biomedical organ anchors and use an organ-wise sparsemax sampler with a learnable abstain anchor to select sparse evidence without fixed Top-k.

2 Related Work

Hallucination-aware RRG methods commonly ground generation through fixed evidence structures. We group them by whether the paradigm is spatial, provided by external vision models, or semantic, specified by entity and phrase vocabularies.

Grounding via external pretrained vision models. Early grounded RRG systems injected evidence from separately trained visual modules into the report decoder. RGRG (Tanida et al., 2023) routes detector-proposed anatomical regions to region-wise report generation, while MAIRA-2 (Bannur et al., 2024) extends grounded reporting by jointly producing findings and bounding boxes. With the rise of universal and medical segmentation models, including TotalSegmentator (Wasserthal et al., 2023), SAT (Zhao et al., 2023), SAM (Kirillov et al., 2023), and MedSAM (Ma et al., 2024), organ- and anatomy-level masks have become accessible external priors for 3D RRG. Recent methods exploit such priors in different forms: Reg2RG (Chen et al., 2025a) uses segmenter-derived masks for local feature decoupling and region–report alignment; CT-GRAPH (Kalisch et al., 2025) organizes organ masks into a hierarchical anatomical graph for cross-organ reasoning; and MedRegion-CT (Kyung et al., 2025) uses anatomical pseudo-masks to guide token pooling across global and local contexts. A related direction replaces explicit masks with vision–language pretraining: fVLM (Shui et al., 2025), for example, aligns organ regions with corresponding report sentences to expose organ-aware features to downstream decoders. These methods improve interpretability, but their grounding quality remains tied to the coverage and calibration of the upstream visual model; missed regions, mislabeled masks, or unsupported anatomy categories can be propagated to the decoder as apparently grounded evidence.

Grounding via fixed entity or phrase queues. A second line avoids dense pixel-level labels by aligning visual evidence to predefined medical concepts. MAIRA-1 (Hyland et al., 2023), RadGenome-ChestCT (Zhang et al., 2024), and CheX (Müller et al., 2024) instantiate this idea in 2D chest X-ray using curated organ and finding vocabularies. Recent 3D extensions refine the semantic paradigm rather than removing it: Dia-LLaMA (Chen et al., 2025b) learns a disease-prototype memory bank as diagnostic anchors; PCRL (Zheng et al., 2024) composes pathological clues from segmented regions, entity types, and report topics; and DEAR (Zhou et al., 2026) decouples medical entity phrases into organ and lesion streams to reduce intra-entity confusion. Phrase-level grounding is attractive because it provides an explicit semantic bottleneck, yet radiological abnormalities are long-tailed and compositional. Fixed vocabularies therefore bias alignment toward frequent or predefined concepts, leaving rare, subtle, or institution-specific findings weakly grounded.

3 HEATRAD

We study *anomaly-guided 3D radiology report generation*: given a volume $\mathbf{V} \in \mathbb{R}^{D\times H\times W}$ with axial slices $\{\mathbf{s}_j\}_{j=1}^{D}$, the model generates a report $\mathbf{y} = (y_1, \ldots, y_T)$ grounded in scan-supported abnormalities. HEATRAD separates *where to look* from *what to say* by routing suspicious slice evidence into an organ-ordered visual prefix $\mathcal{P}$ through suspicious-slice discovery, text-anchored

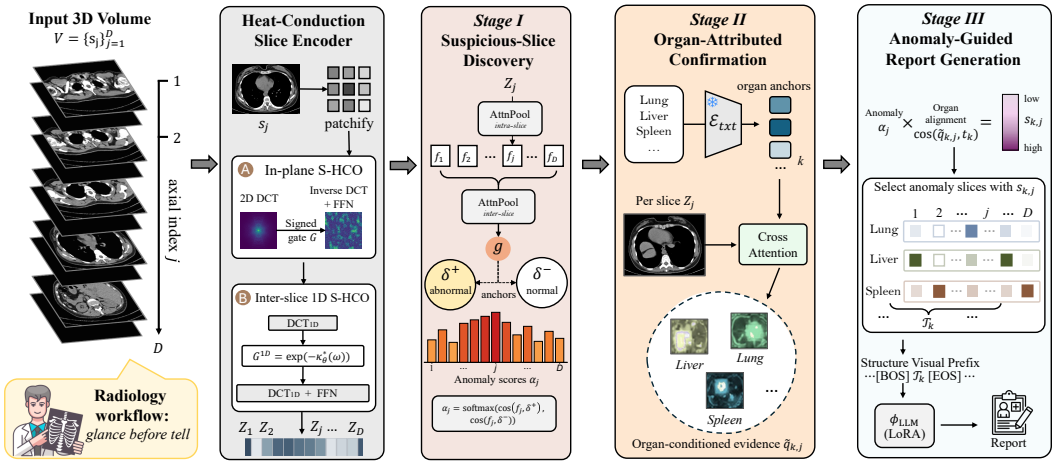

Figure 2: **Overview of HEATRAD.** HEATRAD encodes 3D CT volumes with signed heat-conduction operators, discovers suspicious slices, attributes them to organ contexts, and forms an anomaly-guided visual prefix for report generation.

organ attribution, and sparsemax-based evidence selection. The report is generated autoregressively:

$$p(\mathbf{y} \mid \mathbf{V}) = \prod_{r=1}^{T} p_{\Phi_{\mathrm{LLM}}}\big(y_r \mid y_{<r}, \mathcal{P}(\mathbf{V})\big), \tag{1}$$

where $\mathcal{P}(\mathbf{V})$ is a differentiable prefix of organ-conditioned slice evidence, allowing visual modules to be trained from the report-level objective. Figure 2 illustrates the pipeline.

3.1 Heat-Conduction Slice Encoder

Self-attention (Vaswani et al., 2017) incurs $\mathcal{O}(N^2)$ mixing at volumetric scale and learns dense affinities over CT volumes dominated by smooth anatomy, potentially diluting sparse lesion evidence, while the Heat Conduction Operator (HCO) (Wang et al., 2025) provides $\mathcal{O}(N^{1.5})$ frequency-conditioned mixing. However, direct isotropic 3D HCO, as used in recent segmentation encoders (Wu & Yu, 2025), may smooth localized findings across slices and obscure slice-level lesion evidence. We instead adopt a slice-centric signed HCO (S-HCO), which attenuates anatomical baselines and amplifies lesion-like residues within each slice while keeping axial propagation lightweight and diffusive. The encoder couples an in-plane S-HCO stack with a lightweight inter-slice HCO, enabling axial context propagation while preserving slice-level addressability for anomaly scoring.

In-plane heat conduction. Each slice $\mathbf{s}_j$ is patchified and embedded as $\mathbf{Z}_j^{(0)} \in \mathbb{R}^{H_0 \times W_0 \times C_0}$, then processed by L hierarchical S-HCO stages. At stage ℓ, S-HCO applies a channel-wise spectral filter:

$$\widehat{\mathbf{Z}}_j^{(\ell-1)}(u,v,c) = \mathrm{DCT}\Big[\mathbf{Z}_j^{(\ell-1)}(\cdot,\cdot,c)\Big](u,v), \tag{2}$$

$$\mathbf{G}_\ell(u,v) = \exp\big(-\kappa_{\boldsymbol{\theta}_\ell}(u,v)\big), \tag{3}$$

$$\mathbf{Z}_j^{(\ell)} = \mathrm{DCT}^{-1}\Big[\mathbf{G}_\ell \odot \widehat{\mathbf{Z}}_j^{(\ell-1)}\Big] + \mathrm{FFN}_\ell\Big(\mathrm{Proj}_\ell(\mathbf{Z}_j^{(\ell-1)})\Big). \tag{4}$$

Here DCT / DCT^{-1} are channel-wise 2-D DCT transforms, (u,v) indexes spatial frequency, and Proj_ℓ is the identity or a shape-matching projection. The signed cumulative diffusivity $\kappa_{\boldsymbol{\theta}_\ell} : \mathbb{N}^2 \to \mathbb{R}$ follows the frequency-value-embedding parameterization of HCO (Wang et al., 2025), but removes its positivity constraint: positive values attenuate frequencies, whereas negative values amplify them. This preserves the efficiency of HCO while generalizing diffusion into a signed spectral filter for lesion-aware slice encoding.

Inter-slice heat conduction. After the final in-plane stage, we stack the slice features as $\overline{\mathbf{Z}} \in \mathbb{R}^{D \times H_L \times W_L \times C_L}$ and apply a 1-D HCO Ψ along the slice axis:

$$\mathbf{Z}_j = \Psi(\overline{\mathbf{Z}})_j \in \mathbb{R}^{N_L \times C_L}, \qquad N_L = H_L W_L, \quad j \in [D]. \tag{5}$$

Ψ follows Eqs. 2–4 with a 1-D DCT over slice frequency $\omega \in \{0, \dots, D-1\}$ and a learnable diffusivity $\kappa^*_{\boldsymbol{\theta}_*} : \mathbb{N} \to \mathbb{R}_{>0}$; here we retain the original non-negative form, since the slice axis primarily supplies anatomical continuity and amplification along it would blur slice identities required by Stages I–III.

Our design differs from the prior slice-decoupled encoders. We apply signed HCO only within each slice and use a lightweight diffusive 1-D HCO for late axial context. Unlike attention-based slice decoupling (Ren & Li, 2025; Yan et al., 2022), our encoder provides sub-quadratic spectral mixing with an explicit anatomy-versus-lesion frequency bias, akin to structure-artifact separation in other imaging domains (Peng et al., 2024).

3.2 Stage I: Suspicious-Slice Discovery

Stage I is a differentiable triage module: it converts the encoded volume into slice-level suspiciousness before any organ- or language-level decision is made. This design reduces the downstream search space while avoiding hard Top-k pruning that would prematurely discard subtle findings.

Look-back aggregation. We summarize $\mathbf{Z}_j$ via single-query attentive pooling that adaptively attends to lesion-bearing tokens amplified by S-HCO, and obtain a volume reference by another single-query attentive pooling across slices:

$$\mathbf{f}_j = \mathrm{AttnPool}(\mathbf{Z}_j;\, \mathbf{q}_{\mathrm{slice}}), \qquad \mathbf{g} = \mathrm{AttnPool}(\{\mathbf{f}_j\}_{j=1}^{D};\, \mathbf{q}_{\mathrm{vol}}), \tag{6}$$

where $\mathbf{q}_{\mathrm{slice}}, \mathbf{q}_{\mathrm{vol}} \in \mathbb{R}^{C_L}$ are independent learnable query vectors.

Class-aware anchors. The anchors implement a lightweight normal–abnormal comparator whose decision boundary adapts to the scan-level context. Two learnable prototypes $\boldsymbol{\eta}^+, \boldsymbol{\eta}^- \in \mathbb{R}^{C_L}$ represent abnormality and normality (Xu et al., 2025). They are fused with the global descriptor to form volume-conditioned anchors:

$$\boldsymbol{\delta}^+ = \mathcal{G}^+\big([\boldsymbol{\eta}^+;\, \mathbf{g}]\big), \qquad \boldsymbol{\delta}^- = \mathcal{G}^-\big([\boldsymbol{\eta}^-;\, \mathbf{g}]\big), \tag{7}$$

where $[\cdot;\cdot]$ denotes concatenation and $\mathcal{G}^+, \mathcal{G}^- : \mathbb{R}^{2C_L} \to \mathbb{R}^{C_L}$ are two-layer MLPs with independent parameters.

Slice anomaly score. The score is kept continuous so that later organ attribution can be softly gated rather than discretely filtered. The anomaly probability of slice j is the softmaxed cosine alignment with the two anchors:

$$\alpha_j = \frac{\exp\big(\cos(\mathbf{f}_j,\, \boldsymbol{\delta}^+)/\tau_{\mathrm{s}}\big)}{\exp\big(\cos(\mathbf{f}_j,\, \boldsymbol{\delta}^+)/\tau_{\mathrm{s}}\big) + \exp\big(\cos(\mathbf{f}_j,\, \boldsymbol{\delta}^-)/\tau_{\mathrm{s}}\big)}, \tag{8}$$

where $\tau_{\mathrm{s}} > 0$ is a learnable temperature, parameterized by a softplus transform to ensure positivity. The continuous probability α_j is passed unchanged to Stages II and III, where it serves as a multiplicative gate (Eq. 13); this keeps slice scoring differentiable end-to-end and lets the report objective shape α_j through the downstream gradient path.

3.3 Stage II: Organ-Attributed Confirmation

Stage II grounds suspicious evidence in anatomy before decoding. Directly feeding lesion-like tokens to the LLM forces the decoder to infer organ semantics from local cues and can verbalize weak false positives. Since abnormalities are organ-dependent and organ layout is stable, we bind suspicious evidence to coarse organs, yielding interpretable groups that suppress organ-inconsistent activations.

Text-anchored organ semantics. HEATRAD uses frozen biomedical text anchors as a lightweight anatomical coordinate system, avoiding external mask predictors and voxel-level structures mismatched with slice routing. Let $\mathcal{V}_{\mathrm{org}} = \{w_k\}_{k=1}^{K}$ be fixed coarse organ prompts, e.g., `the lung region`, `the liver region`, and `the spleen region`. Each prompt is encoded once by a frozen text encoder pretrained for anatomy-level CT–report alignment (Shui et al., 2025):

$$\mathbf{t}_k = \mathcal{E}_{\mathrm{txt}}(w_k) \in \mathbb{R}^d, \qquad k \in [K]. \tag{9}$$

Freezing $\mathcal{E}_{\mathrm{txt}}$ keeps the organ anchors stable in this medical-semantic space, allowing visual evidence to be grouped by anatomy without explicit segmentation supervision.

Organ-conditioned slice evidence via cross-attention. This operation turns each organ prompt into a soft spatial probe over each slice, producing evidence tokens that are anatomically indexed but still visually grounded. We use the text anchors $\{\mathbf{t}_k\}$ as cross-attention queries over slice tokens. For each slice $j \in [D]$ and organ $k \in [K]$:

$$\mathbf{U}_j = \mathbf{Z}_j \mathbf{W}_{\text{vis}} \in \mathbb{R}^{N_L \times d}, \tag{10}$$

$$\mathbf{a}_{k,j} = \text{softmax}\Big(\frac{(\mathbf{t}_k \mathbf{W}_Q)(\mathbf{U}_j \mathbf{W}_K)^\top}{\sqrt{d}} \Big) \in \mathbb{R}^{N_L}, \tag{11}$$

$$\widetilde{\mathbf{q}}_{k,j} = \mathbf{a}_{k,j}\, \mathbf{U}_j \mathbf{W}_V \in \mathbb{R}^d, \tag{12}$$

where $\mathbf{W}_{\text{vis}} \in \mathbb{R}^{C_L \times d}$ and $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d \times d}$ are learnable projections. The output $\widetilde{\mathbf{q}}_{k,j}$ is an organ-conditioned slice evidence token: a compact summary of slice j spatially focused on the visual evidence associated with organ prompt w_k.

3.4 Stage III: Anomaly-Guided Report Generation

Stage III is an evidence packaging module for language generation. Rather than passing dense volume tokens to the decoder, it selects organ-specific evidence and arranges it into a structured prefix, so that the report is generated from a compact set of localized visual cues.

Slice–organ relevance score. For each slice j and organ k, we define a scalar relevance score by gating evidence–organ alignment with the slice anomaly probability:

$$s_{k,j} = \alpha_j \, \frac{\cos\big(\widetilde{\mathbf{q}}_{k,j},\, \mathbf{t}_k\big)}{\tau_{\text{o}}}, \qquad (j,k) \in [D] \times [K], \tag{13}$$

where $\tau_{\text{o}} > 0$ is a learnable temperature, parameterized by a softplus transform to ensure positivity. The multiplicative gate α_j makes organ relevance conditional on slice-level suspiciousness, while the cosine term normalizes the evidence–organ alignment.

Adaptive sparsemax slice selection. Reportable findings are sparse and organ-specific slice extents vary across cases, making fixed Top-k selection or dense softmax weighting unsuitable. For each organ k, we apply sparsemax (Martins & Astudillo, 2016) to slice relevance scores augmented with a learnable abstain anchor a_k:

$$\big[\pi_{k,0}, \pi_{k,1}, \ldots, \pi_{k,D}\big] = \text{sparsemax}(a_k, s_{k,1}, \ldots, s_{k,D}), \tag{14}$$

where $j{=}0$ denotes abstention. The resulting weights are sparse over slices, with $\pi_{k,0}$ absorbing mass when no reliable organ-specific evidence is detected. This avoids fixed Top-k constraints while preserving differentiability.

Structured visual prefix. The prefix format exposes both the selected evidence and its anatomical order to the language model. During training, we retain all slice slots and weight them by sparsemax:

$$\bar{\mathbf{q}}_{k,j} = \pi_{k,j}\, \widetilde{\mathbf{q}}_{k,j}, \qquad j = 1, \ldots, D. \tag{15}$$

Zero-weight slots contribute no visual signal but preserve a fixed computation graph. For compact decoding, zero-weight slots can be omitted at inference. The organ-specific evidence bag is

$$\mathcal{T}_k = \big\{ \bar{\mathbf{q}}_{k,j} : j \in [D] \big\}. \tag{16}$$

Concatenating organ bags in canonical anatomical order yields the visual prefix:

$$\mathcal{P} = \Big\Vert_{k=1}^{K} \big[\, \mathbf{d}_k^{\text{beg}};\, \mathcal{T}_k;\, \mathbf{d}_k^{\text{end}} \,\big], \tag{17}$$

where $\|$ denotes ordered token concatenation and $\mathbf{d}_k^{\text{beg}}, \mathbf{d}_k^{\text{end}} \in \mathbb{R}^d$ are organ-specific delimiters obtained from a shared learnable bias plus a projection of $\mathbf{t}_k$. When $\pi_{k,0}$ dominates, the actual slice slots for organ k are near-zero, signaling that the organ was checked but no reliable abnormal evidence was selected.

Autoregressive decoding. A pretrained language model Φ_{LLM} generates the report conditioned on $\mathcal{P}$ with an *instruction*, and the full framework is trained with the autoregressive report loss:

$$\mathcal{L} = -\sum_{r=1}^{T} \log p_{\Phi_{\mathrm{LLM}}}\big(y_r \,\big|\, \mathcal{P},\, y_{<r}\big). \tag{18}$$

During training, each organ bag $[\mathbf{d}_k^{\mathrm{beg}}; \mathcal{T}_k; \mathbf{d}_k^{\mathrm{end}}]$ is independently dropped with probability p_{drop}, annealed from 0.5 to 0. At inference, $p_{\mathrm{drop}} = 0$, and no anomaly labels, organ masks, or lesion annotations are required.

4 Experiment

4.1 Datasets

To ensure fair comparison with recent 3D RRG studies, we conduct all experiments on three chest CT datasets:

RadGenome-ChestCT (Zhang et al., 2024) extends CT-RATE (Hamamci et al., 2024b) with region-level anatomical annotations and provides 25,692 volume–report pairs from 21,304 patients. All volumes are resampled to $1\times1\times3\,\mathrm{mm}^3$ voxels. We follow the official training/evaluation split, using 24,128 pairs for training and 1,564 pairs for evaluation. **BIMCV-R** (Chen et al., 2024), derived from BIMCV-COVID-19+ (Vayá et al., 2020), contains 8,069 chest CT volumes covering 96 disease entities across oncologic, infectious, cardiopulmonary, and respiratory categories. We use a 7:2:1 train/validation/test split. **CTRG-Chest** contains 1,804 paired CT volumes and radiology reports. Following common 3D radiology report generation practice, we retain only the findings section as the language target and use the commonly adopted 7:1:2 train/validation/test split (Deng et al., 2025). All datasets are de-identified. RadGenome-ChestCT follows CT-RATE (CC BY-NC-SA 4.0), BIMCV-R follows the BIMCV research license, and CTRG-Chest is distributed by its authors for research use.

4.2 Implementation Details

Architecture. The slice encoder is a hierarchical S-HCO with a 4×4 patch embedding and channel widths of $\{96, 192, 384, 768\}$. The signed cumulative diffusivity $\kappa_{\boldsymbol{\theta}_\ell}$ follows the frequency-value-embedding parameterization of HCO (Wang et al., 2025), with the original non-negativity constraint removed to allow both attenuation and amplification. The inter-slice operator Ψ keeps the original positive parameterization. The vision–text joint dimension is set to $d{=}768$. In Stage II, we use $K{=}32$ coarse organ prompts drawn from RadLex (Langlotz, 2006). The text encoder $\mathcal{E}_{\mathrm{txt}}$ is instantiated as the language branch of fVLM (Shui et al., 2025) and kept frozen throughout training. In Stage III, each per-organ abstain anchor a_k is initialized to 1.0, making abstention the default sparsemax threshold; the temperatures τ_{s} and τ_{o} are initialized to 0.1 and 0.5. The organ-bag dropout rate p_{drop} is linearly annealed from 0.5 to 0 during the first half of training.

Training. We use Llama3-8B (Grattafiori et al., 2024) as the report decoder Φ_{LLM} for fair comparison, and fine-tune it with LoRA (Hu et al., 2022) of rank $r{=}64$ on the attention and MLP projections. The visual stack and Stages I–III are jointly optimized with the autoregressive objective in Eq. 18. We use AdamW with $(\beta_1, \beta_2){=}(0.9, 0.999)$ and weight decay 0.05. The learning rates are 1×10^{-4} for the visual modules and 2×10^{-4} for LoRA, with a one-epoch linear warm-up followed by cosine decay over 30 epochs. The effective batch size is 16 volumes. A unified CT preprocessing pipeline is applied to all datasets. HU intensities are clipped to $[-1000, 1000]$, z-score normalized using training-split statistics, and resampled along the axial axis to a fixed slice budget of $D{=}64$ slices at 256×256 in-plane resolution. All experiments are implemented in PyTorch $2.5.1$ with CUDA 12.4 and conducted on 8 NVIDIA GeForce RTX 4090 GPUs (24 GB each).

4.3 Evaluation Metrics

Following standard 3D radiology report generation protocols (Hamamci et al., 2024a; Chen et al., 2025a), we report BLEU-1 (B-1), BLEU-4 (B-4) (Papineni et al., 2002), METEOR (MTR) (Banerjee & Lavie, 2005), and ROUGE-L (R-L) (Lin, 2004) for natural language quality. Entity-level

Table 1: Quantitative comparison on three public 3D CT report-generation benchmarks. **Bold** marks the best, underline the second-best. ‡: 2D baselines adapted to 3D inputs via the M3D-CLIP encoder (Bai et al., 2024a); †: results we reproduced under a unified setting when not reported in the original paper; DEAR natively shares the same Llama3-8B report decoder.

Dataset	Method	NLG Metrics				Clinical Efficacy		
		B-1	B-4	MTR	R-L	Prec.	Rec.	F1
RadGenome-ChestCT	FactMM-RAG‡ (ACL'2025)	0.436	0.231	0.390	0.305	0.405	0.142	0.210
	R2GenCSR‡ (JBHI'2026)	0.441	0.234	0.395	0.310	0.410	0.144	0.214
	CT2Rep (MICCAI'2024)	0.444	0.236	0.402	0.310	0.317	0.089	0.139
	RadFM (*Nat Commun*'2025)	0.442	0.237	0.399	0.315	0.382	0.131	0.195
	Reg2RG (TMI'2025)	0.473	0.249	0.441	0.367	0.423	0.181	0.253
	DEAR (AAAI'2026)	0.491	0.366	0.462	0.472	0.782	0.598	0.677
	HeatRAD	**0.508**	**0.388**	**0.481**	**0.491**	**0.805**	**0.621**	**0.702**
BIMCV-R	FactMM-RAG‡ (ACL'2025)	0.395	0.252	0.415	0.460	0.335	0.116	0.174
	R2GenCSR‡ (JBHI'2026)	0.404	0.249	0.418	0.463	0.341	0.119	0.178
	CT2Rep (MICCAI'2024)	0.410	0.285	0.463	0.483	0.326	0.125	0.182
	RadFM† (*Nat Commun*'2025)	0.404	0.278	0.455	0.475	0.319	0.111	0.165
	Reg2RG† (TMI'2025)	0.439	0.304	0.486	0.500	0.351	0.137	0.197
	DEAR (AAAI'2026)	0.447	0.312	0.490	0.519	0.402	0.188	0.256
	HeatRAD	**0.465**	**0.328**	**0.501**	**0.537**	**0.428**	**0.205**	**0.281**
CTRG-Chest	FactMM-RAG‡ (ACL'2025)	0.416	0.302	0.425	0.470	0.295	0.158	0.208
	R2GenCSR‡ (JBHI'2026)	0.420	0.298	0.430	0.473	0.298	0.161	0.213
	CT2Rep (MICCAI'2024)	0.423	0.292	0.470	0.502	0.325	0.102	0.155
	RadFM (*Nat Commun*'2025)	0.487	0.309	0.492	0.491	–	–	–
	Reg2RG (TMI'2025)	0.496	0.320	0.497	0.478	–	–	–
	DEAR (AAAI'2026)	0.502	0.329	0.501	0.514	0.478	0.225	0.306
	HeatRAD	**0.525**	**0.339**	**0.528**	**0.541**	**0.498**	**0.251**	**0.328**

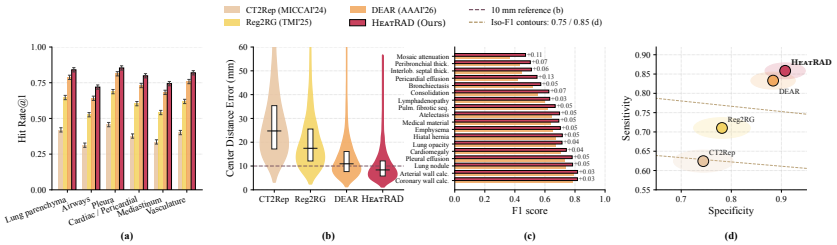

Figure 3: Spatial grounding and patient-level clinical evaluation on the RadGenome-ChestCT. **(a)** Finding-to-region Hit Rate@1 across six anatomical regions. **(b)** Violin plots of CDE between predicted and ground-truth lesion centroids. **(c)** Per-finding F1 of HeatRAD vs. DEAR over the 18 abnormality categories. **(d)** Patient-level Sensitivity vs. Specificity with 95% CI.

clinical consistency is evaluated by extracting clinical entities from generated and reference reports. We use the dataset-provided entity inventories for RadGenome-ChestCT and CTRG-Chest, and RadGraph (Delbrouck et al., 2024) for BIMCV-R. We summarize entity matching with micro-averaged Recall (Rec.), Precision (Prec.), and F1. Complementary clinical fidelity is measured using GREEN (Ostmeier et al., 2024) and RaTEScore (RaTE) (Zhao et al., 2024).

Spatial grounding is evaluated on RadGenome-ChestCT using its region-level anatomical and abnormality annotations. For each organ k, we compute a slice-resolved evidence centroid

$$\hat{\mathbf{c}}_k = \frac{\sum_{j:\pi_{k,j}>0} \pi_{k,j}(\bar{x}_{k,j}, \bar{y}_{k,j}, z_j)}{\sum_{j:\pi_{k,j}>0} \pi_{k,j}}, \tag{19}$$

$\pi_{k,j}$ is the sparsemax-normalized slice weight, $(\bar{x}_{k,j}, \bar{y}_{k,j})$ the in-plane centroid of $\mathbf{a}_{k,j}$, and z_j the axial coordinate. The denominator accounts for abstention, with all-abstain organs counted as misses. We evaluate $\hat{\mathbf{c}}_k$ against organ and associated abnormality regions using Hit Rate and centroid distance error (CDE, mm). For bilateral, multifocal, or diffuse findings, we also report lesion-wise coverage, where an instance is covered if an evidence point falls within its mask or 15 mm, with bilateral organs evaluated per side.

4.4 Comparison with the State-of-the-Art

We compare HEATRAD with representative 3D RRG methods, including CT2Rep (Hamamci et al., 2024a), RadFM (Wu et al., 2025), Reg2RG (Chen et al., 2025a), and DEAR (Zhou et al., 2026). For completeness, we also benchmark two strong 2D RRG baselines, FactMM-RAG (Sun et al., 2025) and R2GenCSR (Wang et al., 2026).

Report-level quality. Tab. 1 compares report quality on three benchmarks. HEATRAD ranks first on all metrics and datasets. F1 improves over DEAR from 0.677 to 0.702 on RadGenome-ChestCT, 0.256 to 0.281 on BIMCV-R, and 0.306 to 0.328 on CTRG-Chest. All HEATRAD results are three-seed means (F1 std ≤ 0.006), and the gains persist against a DEAR-matched control retrained under the identical protocol ($+0.013$/$+0.013$/$+0.011$ F1 on the three datasets; bootstrap 95% CIs exclude zero). NLG gains are smaller, indicating that improvements mainly come from entity-level clinical correctness. The gap to Reg2RG and DEAR supports anomaly-first evidence routing over dense pseudo-mask grounding or fixed phrase queues, while weaker 2D-adapted baselines highlight the need for axial evidence in 3D reporting.

Spatial grounding and patient-level analysis. Fig. 3 shows that HEATRAD improves Hit Rate@1 across six anatomical regions and reduces centroid distance on RadGenome-ChestCT, where the learned slice scores also align with ground-truth abnormalities: against lesion masks, α_j attains slice-level AUROC 0.871 and Recall@8 0.831, versus 0.71/0.44 for a DEAR cross-attention proxy. Lesion-wise coverage is 0.74 overall and 0.69 on multi-lesion volumes. The largest gains over DEAR are observed in long-tail abnormality categories. Patient-level analysis places HEATRAD in the high-sensitivity and high-specificity region under 95% Wilson confidence ellipses, reflecting fewer misses and false positives.

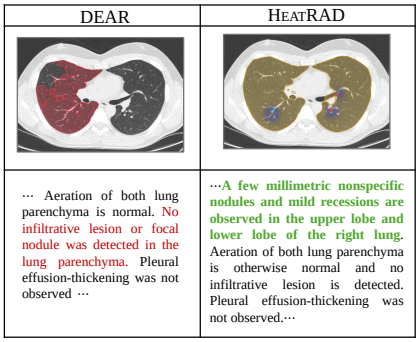

Figure 4: Case study on RadGenome-ChestCT.

Qualitative case. Fig. 4 shows that DEAR partially covers the right lung and misses nodules and recessions with a negated report (red). HEATRAD covers the full lung, selects lesion-supporting slices, and generates a positive report matching the reference (green).

4.5 Ablation Study

Core components. Tab. 2 ablates HEATRAD on RadGenome-ChestCT. Removing S-HCO lowers F1 by 2.1 pp and Hit@1 by 5.2 pp. Stage I is most critical for grounding, dropping Hit@1 to 0.612 without slice anomaly probabilities, while Stage II mainly improves clinical precision, with Prec. falling to 0.782 without organ attribution. Dense pooling in place of the Stage III sparsemax sampler degrades NLG and grounding. Two-stage removals compound the loss, and the vanilla autoregressive baseline performs worst across metrics.

Visual encoder. Tab. 3a compares encoders under the same pipeline. Signed HCSE matches or exceeds signed self-attention using $0.61\times$ parameters. Unsigned HCO lowers F1 by 1.5 pp, supporting signed diffusion. Full 3D HCO slightly improves RaTE but harms Hit@1 ($0.793\rightarrow0.756$), indicating blurred slice evidence. The plain 3D ViT lags despite the largest parameter count.

Anomaly source and sampler. Tab. 3b isolates slice scoring and sampling. Uniform, Random, and Inverted α_j all degrade clinical and grounding metrics, with Inverted lowest on Hit@1 (0.182). Learned α_j attains 96% of Oracle F1 and 79% of Oracle Hit@1 without abnormality supervision. Hard Top-K (K=8) reduces F1 by 1.9 pp and Hit@1 by 3.8 pp, showing the value of adaptive sparse selection.

Table 2: Core-component ablation of HEATRAD on RadGenome-ChestCT. HCSE denotes the Heat-Conduction Slice Encoder. (✗) means: uniform anomaly prior, mean-pooled global evidence bag, dense softmax pooling, or unsigned HCO encoder, with the remaining pipeline unchanged. Hit Rate@1 is averaged over the six anatomical regions in Fig. 3.

Variant	Components				NLG		CE		Clinical		Grnd.
	HCSE	Stage I	Stage II	Stage III	B-4	R-L	Prec.	F1	GREEN	RaTE	Hit@1
HEATRAD	✓	✓	✓	✓	**0.388**	**0.491**	**0.805**	**0.702**	**0.555**	**0.742**	**0.793**
− S-HCO	✗	✓	✓	✓	0.371	0.475	0.785	0.681	0.532	0.727	0.741
− Stage I	✓	✗	✓	✓	0.365	0.470	0.770	0.665	0.521	0.715	0.612
− Stage II	✓	✓	✗	✓	0.378	0.482	0.782	0.679	0.535	0.722	0.685
− Stage III	✓	✓	✓	✗	0.372	0.476	0.752	0.668	0.515	0.712	0.690
− Stages I & II	✓	✗	✗	✓	0.348	0.452	0.730	0.620	0.485	0.682	0.531
− Stages I & III	✓	✗	✓	✗	0.341	0.448	0.715	0.605	0.471	0.671	0.498
− Stages II & III	✓	✓	✗	✗	0.354	0.460	0.738	0.628	0.493	0.690	0.572
Vanilla baseline	✗	✗	✗	✗	0.236	0.310	0.317	0.139	0.180	0.430	0.183

Table 3: More ablation on RadGenome-ChestCT. **(a)** isolates the visual encoder across three orthogonal factors: mixer, polarity, and granularity. For the 3D encoder, the whole-volume evidence collapses Stages I–II into one volumetric anomaly-confirmation block. **(b)** varies only the Stage III slice score α, or replaces sparsemax with tuned hard Top-K sampling (K=8). Score variants are Oracle (radiologist annotations), Uniform (α_j=1), Random ($\alpha_j \sim \mathcal{U}[0,1]$), and Inverted ($1-\alpha_j^*$).

(a) Visual Encoder Probe

Encoder	Params (M)	F1	RaTE	Hit@1
HCSE (Ours)	21.3	**0.702**	0.742	**0.793**
Signed self-attn	34.7	0.696	0.745	0.781
Unsigned HCO	**21.1**	0.687	0.736	0.768
3D HCO	28.5	0.690	**0.748**	0.756
Unsigned self-attn	34.5	0.681	0.730	0.748
3D ViT	86.2	0.674	0.726	0.732

(b) Anomaly Source

Variant	F1	RaTE	Hit@1
Oracle	0.728	0.762	1.000
HEATRAD	**0.702**	**0.742**	**0.793**
Uniform	0.620	0.680	0.380
Random	0.602	0.668	0.265
Inverted	0.480	0.605	0.182
Top-K sampler	0.683	0.725	0.755

5 Conclusion and Limitations

HEATRAD formulates 3D radiology report generation as anomaly-first evidence routing, with a signed heat-conduction slice encoder, organ-attributed confirmation, and sparsemax-based prefix construction trained on volume–report pairs without task-specific spatial supervision. Future work will validate HEATRAD on 3D medical imaging datasets covering additional modalities and anatomical regions, and we commit to using the model strictly as an assistive tool for clinicians. Generated reports used without expert review could mislead clinical decisions, so HEATRAD is intended strictly as a research aid under radiologist supervision.

HEATRAD relies on a frozen biomedical text encoder and fixed RadLex organ prompts, whose errors may propagate under domain shift. Evaluation covers three chest CT benchmarks; other anatomies and demographic subgroups remain untested. The slice suspiciousness score is not a calibrated probability. Post-surgical or congenital anatomy may weaken organ attribution, the 8B decoder dominates inference cost, and no radiologist reader study has been conducted.

Acknowledgement

This work was supported in part by the Major Program of the National Natural Science Foundation of China under Grant 62495064, in part by the Industrial Chain Collaborative Innovation Project of Science and Technology under Grant 2025-XT00-00018-GX, in part by the Sichuan Provincial Medical Imaging Clinical Medicine Research Center Open Project under Grant YXYX2412. Competing interests: The authors declare no competing interests.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 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*, 2024a.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024b.

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.

Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, et al. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024.

Yinda Chen, Che Liu, Xiaoyu Liu, Rossella Arcucci, and Zhiwei Xiong. Bimcv-r: A landmark dataset for 3d ct text-image retrieval. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 124–134. Springer, 2024.

Zhixuan Chen, Yequan Bie, Haibo Jin, and Hao Chen. Large language model with region-guided referring and grounding for ct report generation. *IEEE Transactions on Medical Imaging*, 2025a.

Zhixuan Chen, Luyang Luo, Yequan Bie, and Hao Chen. Dia-llama: Towards large language model-driven ct report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 141–151. Springer, 2025b.

Jean-Benoit Delbrouck, Pierre Chambon, Zhihong Chen, Maya Varma, Andrew Johnston, Louis Blankemeier, Dave Van Veen, Tan Bui, Steven Truong, and Curtis Langlotz. Radgraph-xl: A large-scale expert-annotated dataset for entity and relation extraction from radiology reports. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12902–12915, 2024.

Xiwei Deng, Xianchun He, Jianfeng Bao, Yudan Zhou, Shuhui Cai, Congbo Cai, and Zhong Chen. Mvketr: chest ct report generation with multi-view perception and knowledge enhancement. *IEEE Journal of Biomedical and Health Informatics*, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, 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 *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 476–486. Springer, 2024a.

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, et al. Developing generalist foundation models from a multimodal dataset for 3d computed tomography. *arXiv preprint arXiv:2403.17834*, 2024b.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

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.

Stephanie L Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, et al. Maira-1: A specialised large multimodal model for radiology report generation. *arXiv preprint arXiv:2311.13668*, 2023.

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

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Sunggu Kyung, Jinyoung Seo, Hyunseok Lim, Dongyeong Kim, Hyungbin Park, Jimin Sung, Jihyun Kim, Wooyoung Jo, Yoojin Nam, and Namkug Kim. Medregion-ct: region-focused multimodal llm for comprehensive 3d ct report generation. *arXiv preprint arXiv:2506.23102*, 2025.

Curtis P Langlotz. Radlex: a new method for indexing online educational materials, 2006.

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.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023b.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature communications*, 15(1):654, 2024.

Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International conference on machine learning*, pp. 1614–1623. PMLR, 2016.

Philip Müller, Georgios Kaissis, and Daniel Rueckert. Chex: Interactive localization and region description in chest x-rays. In *European Conference on Computer Vision*, pp. 92–111. Springer, 2024.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, et al. Green: Generative radiology report evaluation and error notation. *arXiv preprint arXiv:2405.03595*, 2024.

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.

Hao Peng, Ting-Zhu Huang, Xi-Le Zhao, Jie Lin, Wei-Hao Wu, and Li-Yuan Li. Deep domain fidelity and low-rank tensor ring regularization for thick cloud removal of multitemporal remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024.

Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. Ai in health and medicine. *Nature medicine*, 28(1):31–38, 2022.

Sucheng Ren and Xiaomeng Li. Hresformer: Hybrid residual transformer for volumetric medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 36(6): 10558–10566, 2025.

Zhongyi Shui, Jianpeng Zhang, Weiwei Cao, Sinuo Wang, Ruizhe Guo, Le Lu, Lin Yang, Xianghua Ye, Tingbo Liang, Qi Zhang, et al. Large-scale and fine-grained vision-language pre-training for enhanced ct image understanding. *arXiv preprint arXiv:2501.14548*, 2025.

Liwen Sun, James Jialun Zhao, Wenjing Han, and Chenyan Xiong. Fact-aware multimodal retrieval augmentation for accurate medical radiology report generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 643–655, 2025.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7433–7442, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Maria De La Iglesia Vayá, Jose Manuel Saborit, Joaquim Angel Montell, Antonio Pertusa, Aurelia Bustos, Miguel Cazorla, Joaquin Galant, Xavier Barber, Domingo Orozco-Beltrán, Francisco García-García, et al. Bimcv covid-19+: A large annotated dataset of rx and ct images from covid-19 patients. *arXiv preprint arXiv:2006.01174*, 2020.

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*, pp. 1–14, 2026. doi: 10.1109/JBHI.2026.3669539.

Zhaozhi Wang, Yue Liu, Yunjie Tian, Yunfan Liu, Yaowei Wang, and Qixiang Ye. Building vision models upon heat conduction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9707–9717, 2025.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5 (5):e230024, 2023.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):7866, 2025.

Rong Wu and Yim-Sang Yu. Enhancing medical image segmentation via heat conduction equation. *arXiv preprint arXiv:2511.03260*, 2025.

Jiacong Xu, Shao-Yuan Lo, Bardia Safaei, Vishal M Patel, and Isht Dwivedi. Towards zero-shot anomaly detection and reasoning with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20370–20382, 2025.

Xiangyi Yan, Hao Tang, Shanlin Sun, Haoyu Ma, Deying Kong, and Xiaohui Xie. After-unet: Axial fusion transformer unet for medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 3971–3981, 2022.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Jiayu Lei, Ya Zhang, Yanfeng Wang, and Weidi Xie. Radgenome-chest ct: A grounded vision-language dataset for chest ct analysis. *arXiv preprint arXiv:2404.16754*, 2024.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Ratescore: A metric for radiology report generation. *arXiv preprint arXiv:2406.16845*, 2024.

Ziheng Zhao, Yao Zhang, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. One model to rule them all: Towards universal segmentation for medical images with text prompts. *arXiv e-prints*, pp. arXiv–2312, 2023.

Chengxin Zheng, Junzhong Ji, Yanzhao Shi, Xiaodan Zhang, and Liangqiong Qu. See detail say clear: Towards brain ct report generation via pathological clue-driven representation learning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 16542–16552, 2024.

Lingyu Zhou, Yue Yu, Zhang Yi, and Xiuyuan Xu. Mitigating entity hallucinations in 3d radiology report generation via dual-stream alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13719–13727, 2026.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction enumerate all concrete contributions.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Sec. 5 discusses the limitations of the method, the grounding metrics, and the evaluation scope.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The contribution is algorithmic and architectural; no formal theorems or theoretical guarantees are claimed.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The Method section fully specifies the architecture, the training objective, and the data flow. The Experiment section reports the optimizer, schedule, batch size, LoRA rank, decoding hyper-parameters, and dataset splits. Training and evaluation scripts are publicly available at `https://github.com/SCU-zly/HeatRAD`.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code is publicly available at `https://github.com/SCU-zly/HeatRAD`, including training, inference, and pre-processing scripts together with a model card. All datasets used are publicly available under their original licenses.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Experiment section reports all the training and test details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Main results are means over three training seeds, and the matched-control comparison is supported by paired bootstrap confidence intervals; the patient-level analysis (Fig. 3d) reports 95% Wilson confidence intervals; remaining results are evaluated on fixed official splits.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The Experiment section reports the GPU type, count, and memory (8× RTX 4090, 24 GB each) and the software environment (PyTorch 2.5.1, CUDA 12.4).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses publicly released, de-identified medical-imaging datasets under their original licenses, makes no claim of clinical deployment readiness, and is intended as research toward safer, more grounded automatic radiology reporting. The submission preserves anonymity per double-blind requirements.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Conclusion section discusses the social impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The eventual model release will be research-use-only, we do not release any new patient data; only model weights and processing scripts will be made available.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Dataset licenses are stated in the Datasets section; all third-party assets are cited at first use and used under their original terms.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The trained HEATRAD checkpoint, training/evaluation scripts, and preprocessing pipeline will be released after the review period together with a model card and a README documenting expected inputs, outputs, intended use, and known limitations. No new dataset is released.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or any new research with human subjects. All evaluations are conducted on previously released, de-identified datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: This work does not collect new human-subject data. All datasets used were previously released by their original providers under appropriate ethical review; we abide by their data-use agreements.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are used only as a standard pretrained backbone (Llama3-8B) for report decoding; they are not an important, original, or non-standard component of the core methodology.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.