Concept Planning for Structured Chest X-ray Report Generation

Anonymous ACL submission

Abstract

Medical multimodal large language models (MLLMs) have achieved strong results in radiology report generation, producing fluent and clinically informative chest X-ray reports. These models may already possess a latent capacity to anticipate upcoming clinical concepts before verbalizing findings. Nevertheless, how to systematically profile, enhance, and leverage this capacity for faithful report generation remains unclear. To assess it, we introduce the Clinical Concept Anticipation Score (CCAS), a rank-based metric that measures whether upcoming positive clinical concepts are pre-captured in the model's vocabulary distribution before corresponding findings are generated. Our analysis shows that supervised fine-tuned MLLMs exhibit a concept-level anticipatory signal, but this signal remains implicit and weakly optimized by next-token prediction. Based on this observation, we propose ConPlan, a plug-and-play clinical concept planning framework that reformulates reporting as a structured multi-step process with regional finding generation and final synthesis. ConPlan further introduces dual-level planning, where chunk-level prediction anticipates near-future clinical concepts and step-level guidance reuses self-generated intermediate states as priors. It is trained with next-token generation and planning objectives, requiring no external clinical labels during inference. Experiments across multiple backbones and out-of-domain datasets show improvements in report quality, clinical correctness, and concept anticipation, demonstrating the value of explicitly activating clinical concept planning in medical MLLMs.

1 Introduction

Radiology report generation (RRG) translates a chest X-ray study, together with its clinical context, into a faithful diagnostic report (Seah et al., 2021; Rajpurkar et al., 2018; Cid et al., 2024; Wang et al., 2025b). Early RRG systems augmented encoder–decoder backbones with medical memory modules, prior/posterior knowledge injection, expert tokens, and region-level grounding (Chen et al., 2020, 2021; Liu et al., 2021; Wang et al., 2023a; Tanida et al., 2023). The field has since shifted toward medical multimodal large language models (MLLMs), in which a vision encoder is aligned with an LLM so that radiology reporting, medical dialogue, and broader biomedical reasoning can be cast as a single generalist vision-language generation task (Wang et al., 2023b; Liu et al., 2024; Pellegrini et al., 2023; Hyland et al., 2023; Li et al., 2023; Moor et al., 2023; Liu et al., 2026b). Despite this architecture leap, almost all such systems are still optimized by a single, purely local objective: autoregressive next-token prediction (Liu et al., 2026a).

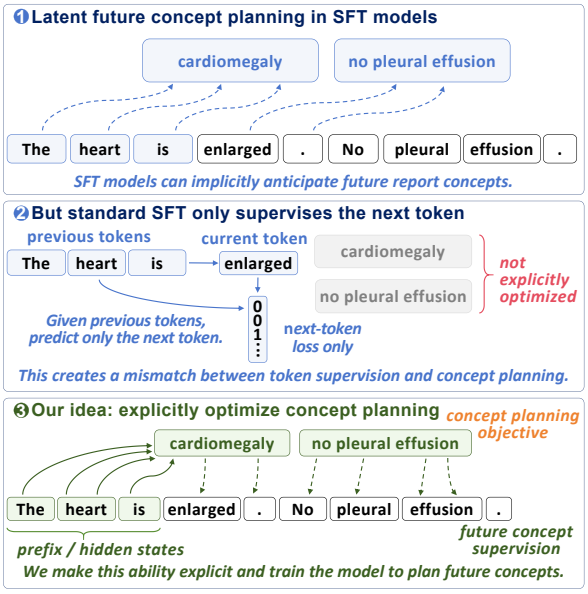


Figure 1: Overview of ConPlan. We first probe latent clinical concept anticipation in an SFT medical MLLM, and then convert this signal into explicit guidance through structured multi-step report generation and dual-level concept planning.

A faithful report, however, is not a long sequence

of locally plausible tokens. It is a structured clinical argument that the model must select the findings that matter, assign each to the correct anatomical or functional region, preserve temporal comparisons against priors, and avoid statements that contradict one another. Region-guided and diagnosis-chain methods have shown that imposing such structure is beneficial (Tanida et al., 2023; Gu et al., 2025; Jin et al., 2025), yet standard MLLM fine-tuning never explicitly ask the decoder to organize upcoming clinical concepts before it begins to verbalize them. The result is a subtle but consequential failure mode. Reports that read fluently but silently omit findings, blur regional boundaries, or drift into clinically inconsistent claims.

Does this planning ability already exist, hidden inside an supervised fine-tuning (SFT) model? To answer this, we probe the decoder before it generates any regional finding. We split each report into five anatomical/functional regions (lungs, heart/mediastinum, pleura, bones/soft tissues, devices/lines and tubes), and at the very first decoding step of each region we look at the next-token distribution–before a single finding word has been emitted. We then introduce the Clinical Concept Anticipation Score (CCAS), a rank-based probe that asks a simple question: are the positive clinical concepts that should appear next already pushed toward the top of the vocabulary, relative to clinically plausible negatives? The answer is yes but only weakly. Positive concepts reach an MRR of 0.0167 versus 0.0069 for negatives, and a top-500 recall of 70.09% versus 34.59%. Next-token supervision therefore does not leave the model purely myopic; it deposits a faint but real anticipatory signal. Because this signal is a by-product of token-level supervision rather than its target, it is uncalibrated, unstable, and never directly optimized.

In this paper, our central proposal is to convert this latent signal into an explicit training objective. We present **ConPlan**, a structured clinical concept-planning framework that turns end-to-end image-to-report generation into a plan-then-write process. The model first produces region-specific findings (*i.e.*, lungs, heart/mediastinum, pleura, bones/soft tissues, devices/lines and tubes), and only then synthesizes a final, coherent report. This decomposition exposes clinically meaningful intermediate states, makes each generation step locally auditable, and aligns supercision with the way radiologists actually reason about an image.

On top of this multi-step backbone we install a dual-level planner inside an intermediate decoder layer. (1) The chunk-level brance groups consecutive hidden states into short chunks and projects them onto a discrete clinical-concept space, training the decoder to anticipate near-future positive concepts before they are verbalized. (2) The step level brance reuses the model's self-generated regional clinical state as an internal prior, conditioning the corresponding report span on a concept summary the model itself has just produced.

The contributions of this work are summarized as follows:

- We identify and quantify implicit clinical concept anticipation in SFT medical MLLMs through CCAS, showing that upcoming positive concepts are already more visible than clinical negatives at the very moment a region begins to be generated.
- We reformulate chest X-ray reporting as a structured multi-step process that produces region-specific findings before synthesizing the final report, providing clearer intermediate supervixion and controllablity.
- We propose ConPlan, a dual-level planner that combines chunk-level future-concept anticipation with step-level self-guided clinical priors, turning concept planning from a by-product into a first-class training signal.
- Experiments and analyses show that ConPlan consistently improves both report quality and clinical correctness, and that it strengthens the anticipatory signal itself, without requiring any external clinical labels at inference.

2 Preliminary

2.1 Regional Stepwise Generation

To make report generation more structured, we introduce **Regional Stepwise Generation** (RSG), which reformulates single-turn chest X-ray report generation into a region-to-report process. For each training sample, we use Qwen3-VL-8B-Instruct as an automatic parser to decompose the target findings into five anatomical or functional regions: lungs, heart/mediastinum, pleura, bones/soft tissues, and devices/lines and tubes. The sample is then converted into a multi-step instruction-following sequence, where the model generates region-specific findings in a fixed order and finally synthesizes them into a complete findings

section. The final-step target remains the original findings, keeping RSG aligned with the standard report-generation objective. The data templates are provided in Appendix.

2.2 Clinical Concept Anticipation Score

Although SFT optimizes only next-token prediction, it remains unclear whether the resulting model has learned any preference for clinical concepts that will appear later in a regional finding. We study this question by decomposing each report into five major anatomical/sectional regions and generating the corresponding findings sequentially. At the start of each region, before any finding text is produced, we inspect the model's next-token distribution and measure whether the positive clinical concepts of that region are already ranked above clinical negatives.

For a study i and region r, let

$$\mathbf{C}_{i,r} = [\mathbf{I}_i, \mathbf{P}_i, \mathbf{Y}_{i,<r}, \mathbf{q}_r] \tag{1}$$

be the context at the start of region r, where $\mathbf{I}_i$ is the image, $\mathbf{P}_i$ is the instruction, $\mathbf{Y}_{i,<r}$ denotes previous regional findings, and $\mathbf{q}_r$ is the current regional prompt. The model then defines a vocabulary distribution $p_{i,r} = P_\theta(\cdot \mid \mathbf{C}_{i,r})$ before generating the first token of the regional finding.

Let $\mathcal{G}_{i,r}$ be the set of positive clinical concepts assigned to region r. Each concept c is mapped to a canonical phrase, whose first token is denoted by $\nu(c)$. We measure its visibility in the step-start distribution by the rank

$$\rho_{i,r}(c) = 1 + \sum_{v \in V} \mathbb{1}\left[p_{i,r}(v) > p_{i,r}(\nu(c))\right]. \tag{2}$$

A smaller $\rho_{i,r}(c)$ means that the positive concept is placed closer to the top of the model's next-token distribution. When a region contains multiple positive concepts, we evaluate all of them rather than selecting the best-ranked one.

To test whether the observed preference is tied to the required findings rather than to medical vocabulary frequency, we compare positive concepts with clinical negatives. For each region, the negative set $\mathcal{N}_{i,r}$ is formed from the same 14-label chest X-ray concept vocabulary after removing the region's positive concepts. We further discard negatives whose first token overlaps with any positive concept, avoiding ambiguous comparisons such as *Lung Lesion* and *Lung Opacity*. Let $\mathcal{A}_{i,r}$ denote either $\mathcal{G}_{i,r}$ or $\mathcal{N}_{i,r}$, and collect all evaluated concepts

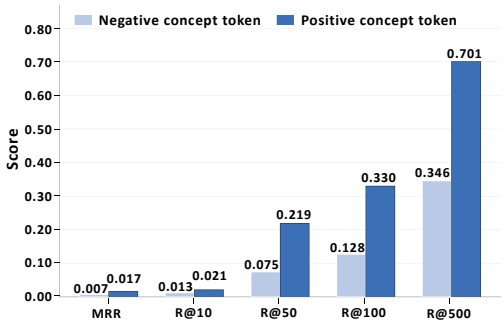


Figure 2: Region-start clinical concept anticipation. Positive concept tokens are ranked higher than clinical negative tokens in the next-token distribution, showing that the SFT model already develops an implicit preference for concepts required by the upcoming regional finding.

as

$$\Omega_{\mathcal{A}} = \{(i, r, c) \mid c \in \mathcal{A}_{i,r}\}. \tag{3}$$

Based on the anticipatory ranks in Eq. 2, we define the **Clinical Concept Anticipation Score** (CCAS) as a rank-based probing score:

$$\mathcal{S}_\phi(\mathcal{A}) = \mathbb{E}_{(i,r,c)\in\Omega_{\mathcal{A}}}\left[\phi(\rho_{i,r}(c))\right], \tag{4}$$

where ϕ maps a concept rank to a utility value. Different choices of ϕ recover the metrics used in our analysis. For recall (Manning et al., 2008) at cutoff m, we set

$$\phi_m^{\mathrm{rec}}(z) = \mathbb{1}[z \le m], \quad m \in \{10, 50, 100, 500\}, \tag{5}$$

which gives $\mathrm{R@}m = \mathcal{S}_{\phi_m^{\mathrm{rec}}}(\mathcal{A})$. For mean reciprocal rank (Voorhees, 1999), we set

$$\phi^{\mathrm{mrr}}(z) = z^{-1}, \tag{6}$$

which gives $\mathrm{MRR} = \mathcal{S}_{\phi^{\mathrm{mrr}}}(\mathcal{A})$.

Fig. 2 highlights two observations. First, positive clinical concepts are already more visible in the model's region-start distribution than negative concepts: the MRR rises from 0.0069 to 0.0167, indicating that the model tends to place positive concepts at higher vocabulary ranks. Second, this advantage persists across different levels of candidate coverage. Within the top-50, top-100, and top-500 predictions, positive concepts are recalled at 21.93%, 33.01%, and 70.09%, respectively, whereas negative concepts reach only 7.47%, 12.75%, and 34.59%. Thus, SFT already induces a certain degree of clinical concept planning before regional finding generation. Yet this ability is only implicitly acquired through next-token learning, rather than directly encouraged by the training objective. Our method therefore turns this latent planning capacity into an explicit training signal.

3 Method

3.1 Overview

We introduce **ConPlan**, a plug-and-play framework that enables clinical concept planning for MLLMs in chest X-ray report generation. As shown in Fig. 3, ConPlan consists of two components: **Regional Stepwise Generation** (RSG) and a dual-level concept planning module. RSG reformulates single-turn image-to-report generation into a multi-step process, where the model first produces findings for five anatomical or functional regions and then synthesizes the final report. The planning module inserts a lightweight concept planning module into an intermediate language decoder layer of a base MLLM. Given the hidden states from the ℓ-th decoder layer,

$$\mathbf{H}^{\ell} = [\mathbf{h}_1^{\ell}, \ldots, \mathbf{h}_T^{\ell}], \quad \mathbf{h}_i^{\ell} \in \mathbb{R}^d, \tag{7}$$

the module constructs concept-level residuals and injects them back into the decoder hidden states. As illustrated in Fig. 3, it contains two complementary branches: a **Chunk-level Concept Planner**, which anticipates near-future concepts from previous token chunks, and a **Step-level Concept Planner**, which uses the model-generated clinical state as a local prior for the corresponding report span. The **Chunk-level Concept Planner** uses Vector Quantization, a Concept Prediction Transformer, and Logit-Guided Codebook Readout to anticipate near-future chunk concepts, whereas the **Step-level Concept Planner** extracts a self-generated intermediate state within each regional step and converts it into step-level concept logits through Logit-Guided Readout. ConPlan is trained with next-token prediction and auxiliary concept planning objectives, and follows the same RSG protocol during inference without requiring external clinical labels.

3.2 Dual-level Clinical Concept Planning

3.2.1 Chunk-level Concept Planner

We first divide the token sequence into non-overlapping chunks of size C. Let $\mathcal{I}_k$ denote the token indices in the k-th chunk. The chunk representation is obtained by average pooling:

$$\mathbf{c}_k = \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \mathbf{h}_i^{\ell}, \quad k = 1, \ldots, K. \tag{8}$$

When the sequence length is not divisible by C, zero padding is applied before pooling.

To form a discrete concept space, each chunk representation is split into S segments,

$$\mathbf{c}_k = [\mathbf{c}_{k,1}; \ldots; \mathbf{c}_{k,S}], \tag{9}$$

where S defaults to the number of text attention heads. For each segment s, we maintain a codebook $\mathbf{E}_s = \{\mathbf{e}_{s,m}\}_{m=1}^{M}$. After a small codebook transformation $\phi(\cdot)$, Vector Quantization assigns each segment to its nearest code:

$$q_{k,s} = \arg\min_{m} \|\mathbf{c}_{k,s} - \phi(\mathbf{e}_{s,m})\|_2^2, \tag{10}$$

and the quantized chunk concept is

$$\bar{\mathbf{c}}_k = [\phi(\mathbf{e}_{1,q_{k,1}}); \ldots; \phi(\mathbf{e}_{S,q_{k,S}})]. \tag{11}$$

A Concept Prediction Transformer then predicts the concept code of each next chunk from previous chunk states. With a causal mask, we compute

$$[\mathbf{z}_1, \ldots, \mathbf{z}_{K-1}] = \mathrm{CPT}(\mathbf{c}_1, \ldots, \mathbf{c}_{K-1}), \tag{12}$$

where $\mathbf{z}_{k-1}$ is used to predict the code distribution of chunk k. For each segment,

$$\mathbf{o}_{k,s}^{\mathrm{ch}} = \mathbf{W}_s^{\mathrm{ch}} \mathbf{z}_{k-1}, \quad k = 2, \ldots, K. \tag{13}$$

Instead of selecting a hard code, ConPlan uses Logit-Guided Codebook Readout to obtain a differentiable concept representation:

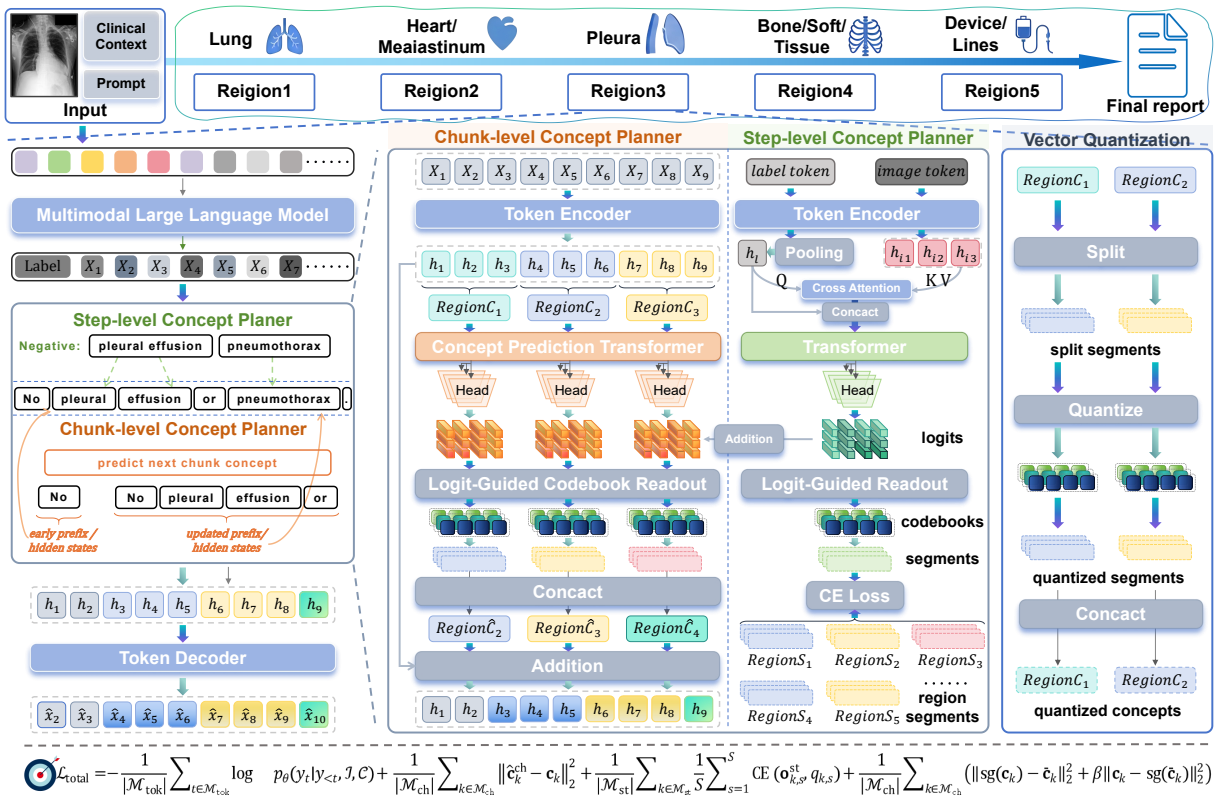
$$\begin{aligned} \boldsymbol{\pi}_{k,s}^{\mathrm{ch}} &= \mathrm{softmax}(\mathbf{o}_{k,s}^{\mathrm{ch}}), \\ \hat{\mathbf{c}}_{k,s}^{\mathrm{ch}} &= \sum_{m=1}^{M} \pi_{k,s,m}^{\mathrm{ch}} \phi(\mathbf{e}_{s,m}), \\ \hat{\mathbf{c}}_k^{\mathrm{ch}} &= [\hat{\mathbf{c}}_{k,1}^{\mathrm{ch}}; \ldots; \hat{\mathbf{c}}_{k,S}^{\mathrm{ch}}]. \end{aligned} \tag{14}$$

The first chunk has no previous chunk-level context, so its chunk-level concept residual is set to zero.

3.2.2 Step-level Concept Planner

The chunk-level planner captures local concept transitions, while each RSG step also provides a step-wise planning state associated with the regional finding. For the r-th valid step, let $\mathcal{S}_r$ and $\mathcal{F}_r$ denote the token positions of the planning-state span and the corresponding finding span, respectively. Empty or placeholder spans are ignored. The planning-state representation is obtained by pooling its hidden states:

$$\mathbf{s}_r = \frac{1}{|\mathcal{S}_r|} \sum_{i \in \mathcal{S}_r} \mathbf{h}_i^{\ell}. \tag{15}$$



$$\mathcal{L}_{\text{total}} = -\frac{1}{|\mathcal{M}_{\text{tok}}|}\sum_{t\in\mathcal{M}_{\text{tok}}} \log\ p_\theta(y_t|y_{<t},\mathcal{I},\mathcal{C}) + \frac{1}{|\mathcal{M}_{\text{ch}}|}\sum_{k\in\mathcal{M}_{\text{ch}}} \|\hat{\mathbf{c}}_k^{\text{ch}} - \mathbf{c}_k\|_2^2 + \frac{1}{|\mathcal{M}_{\text{st}}|}\sum_{k\in\mathcal{M}_{\text{st}}} \frac{1}{S}\sum_{s=1}^{S} \text{CE}\,(\mathbf{o}_{k,s}^{\text{st}}, q_{k,s}) + \frac{1}{|\mathcal{M}_{\text{ch}}|}\sum_{k\in\mathcal{M}_{\text{ch}}} (\|\text{sg}(\mathbf{c}_k) - \hat{\mathbf{c}}_k\|_2^2 + \beta\|\mathbf{c}_k - \text{sg}(\hat{\mathbf{c}}_k)\|_2^2)$$

Figure 3: Overview of **ConPlan**. The framework based on Regional Stepwise Generation. The Chunk-level Concept Planner anticipates near-future chunk concepts in a segmented codebook space, while the Step-level Concept Planner uses model-generated clinical states as self-guided priors for regional report spans. Their logits are fused through logit-guided codebook readout and injected into decoder hidden states for concept-aware report generation.

The planning state is generated within the same autoregressive sequence and is used only through hidden representations. During training, it is observed under teacher forcing; during inference, the module reuses the state tokens already produced in the current step to guide the following finding tokens, without relying on external clinical labels.

To ground the planning state in visual evidence, we collect image-token hidden states $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_N]$. If $N = 0$, we set $\mathbf{a}_r = \mathbf{0}$. Otherwise, we compute

$$\begin{aligned} \mathbf{q}_r = \mathbf{W}_q\mathbf{s}_r, \quad \mathbf{K}_V = \mathbf{W}_k\mathbf{V}, \quad \mathbf{Z}_V = \mathbf{W}_v\mathbf{V}, \\ \mathbf{a}_r = \text{softmax}\left(\frac{\mathbf{q}_r\mathbf{K}_V^\top}{\sqrt{d}}\right)\mathbf{Z}_V. \end{aligned} \tag{16}$$

The state representation and visual context are fused and projected into step-level logits over the shared segmented codebook:

$$\begin{aligned} \mathbf{u}_r &= f_{\text{fuse}}([\mathbf{s}_r; \mathbf{a}_r]), \\ \mathbf{o}_{r,j}^{\text{st}} &= \mathbf{W}_j^{\text{st}}\mathbf{u}_r, \quad j = 1, \ldots, S. \end{aligned} \tag{17}$$

The step-level prior is assigned to chunks overlapping with the associated finding span. Let $m_{r,k} = 1$ if chunk k overlaps with $\mathcal{F}_r$, and $m_{r,k} = 0$ otherwise. For chunks not covered by any valid step, we set $\mathbf{o}_{k,j}^{\text{st}} = \mathbf{0}$. Otherwise, with $n_k = \sum_r m_{r,k}$, the chunk-wise step logits are computed as:

$$\mathbf{o}_{k,j}^{\text{st}} = \frac{1}{n_k}\sum_r m_{r,k}\mathbf{o}_{r,j}^{\text{st}}. \tag{18}$$

3.2.3 Logit Fusion and Concept Injection

The chunk-level anticipation logits and step-level prior logits are fused in the shared codebook space:

$$\mathbf{o}_{k,s}^{\text{fuse}} = \mathbf{o}_{k,s}^{\text{ch}} + \lambda_{\text{st}}\mathbf{o}_{k,s}^{\text{st}}, \quad k = 2, \ldots, K. \tag{19}$$

The fused logits are converted into a soft concept through the same logit-guided readout:

$$\begin{aligned} \boldsymbol{\pi}_{k,s}^{\text{fuse}} &= \text{softmax}(\mathbf{o}_{k,s}^{\text{fuse}}), \\ \hat{\mathbf{c}}_{k,s}^{\text{fuse}} &= \sum_{m=1}^{M} \pi_{k,s,m}^{\text{fuse}}\phi(\mathbf{e}_{s,m}), \\ \hat{\mathbf{c}}_k^{\text{fuse}} &= [\hat{\mathbf{c}}_{k,1}^{\text{fuse}}; \ldots; \hat{\mathbf{c}}_{k,S}^{\text{fuse}}]. \end{aligned} \tag{20}$$

Finally, the fused concept is broadcast to token positions in the corresponding chunk and added to

the decoder hidden states:

$$\tilde{\mathbf{h}}_i^\ell = \mathbf{h}_i^\ell + \Delta\mathbf{h}_i^\ell, \quad \Delta\mathbf{h}_i^\ell = \begin{cases} \hat{\mathbf{c}}_{\kappa(i)}^{\text{fuse}}, & \kappa(i) > 1, \\ \mathbf{0}, & \kappa(i) = 1, \end{cases} \tag{21}$$

where $\kappa(i)$ maps token i to its chunk index. In implementation, an optional one-token shift can be enabled so that the broadcast concept is aligned with the next-token prediction target.

The concept module is trained with three auxiliary objectives. The chunk-level prediction loss encourages the predicted concept to match the next chunk representation:

$$\mathcal{L}_{\text{ch}} = \frac{1}{|\mathcal{M}_{\text{ch}}|} \sum_{k \in \mathcal{M}_{\text{ch}}} \left\| \hat{\mathbf{c}}_k^{\text{ch}} - \mathbf{c}_k \right\|_2^2. \tag{22}$$

The VQ loss regularizes the discrete concept space:

$$\mathcal{L}_{\text{vq}} = \frac{1}{|\mathcal{M}_{\text{vq}}|} \sum_{k \in \mathcal{M}_{\text{vq}}} \big(\|\text{sg}(\mathbf{c}_k) - \bar{\mathbf{c}}_k\|_2^2 + \beta \|\mathbf{c}_k - \text{sg}(\bar{\mathbf{c}}_k)\|_2^2 \big), \tag{23}$$

where $\text{sg}(\cdot)$ denotes stop-gradient. The step-level loss trains the step prior to predict the quantized code indices of finding chunks covered by valid state-to-finding span pairs:

$$\mathcal{L}_{\text{st}} = \frac{1}{|\mathcal{M}_{\text{st}}| S} \sum_{k \in \mathcal{M}_{\text{st}}} \sum_{s=1}^{S} \text{CE}\left(\mathbf{o}_{k,s}^{\text{st}}, q_{k,s}\right). \tag{24}$$

Together with the standard next-token prediction loss, these objectives encourage the base MLLM to anticipate near-future chunk concepts while using its own generated clinical states as step-level priors.

3.3 Training and Inference

Training. We fine-tune MLLM under the standard supervised report generation setting. Given an input image $\mathcal{I}$, clinical context $\mathcal{C}$, and the structured multi-step target sequence $\mathbf{Y}$, the model is trained with the next-token prediction loss over assistant-side tokens:

$$\mathcal{L}_{\text{ntp}} = -\frac{1}{|\mathcal{M}|} \sum_{t \in \mathcal{M}} \log p_\theta(y_t \mid y_{<t}, \mathcal{I}, \mathcal{C}), \tag{25}$$

where $\mathcal{M}$ denotes the supervised token positions. User instruction tokens are ignored in the loss.

The proposed concept planning module is inserted into an intermediate decoder layer of MLLM. During training, it receives the hidden states from this layer and produces concept-aware representations that are injected back into the decoder. The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{ntp}} + \lambda_{\text{ch}} \mathcal{L}_{\text{ch}} + \lambda_{\text{vq}} \mathcal{L}_{\text{vq}} + \lambda_{\text{st}} \mathcal{L}_{\text{st}}. \tag{26}$$

Here, $\mathcal{L}_{\text{ch}}$ supervises chunk-level concept anticipation, $\mathcal{L}_{\text{vq}}$ regularizes the discrete concept codebook, and $\mathcal{L}_{\text{st}}$ trains the step-level clinical prior. The step-level prior is constructed from the intermediate state tokens preceding each regional finding and is trained to predict the concept codes of the associated finding span.

Inference. During inference, we follow the same structured multi-step generation protocol as in training. The model first generates section-level findings for lungs, heart/mediastinum, pleura, bones/-soft tissues, and devices/lines and tubes, and then produces a final integrated report. The whole process can be written as

$$p_\theta(\mathbf{Y} \mid \mathcal{I}, \mathcal{C}) = \prod_{j=1}^{6} p_\theta(\mathbf{Y}^{(j)} \mid \mathcal{I}, \mathcal{C}, \mathbf{Y}^{(<j)}), \tag{27}$$

where $\mathbf{Y}^{(j)}$ denotes the output of the j-th step.

4 Experiments

4.1 Datasets and Evaluation Metrics

We conduct experiments on the MIMIC-CXR dataset (Johnson et al., 2019). We further evaluate the generalization ability of our method on two out-of-domain validation datasets, IU-Xray (Demner-Fushman et al., 2016) and CheXpert Plus (Chambon et al., 2024). We report BLEU-1/4 (Papineni et al., 2002) and ROUGE-L (Lin, 2004); for clinical evaluation, we report CheXbert similarity (Smit et al., 2020; Yu et al., 2023), F1-CheXbert over 14 thoracic observations and 5 key observations (Irvin et al., 2019; Smit et al., 2020), RadGraph-F1 (Jain et al., 2021), RadCliQ_0 (Yu et al., 2023), and RG_{ER} (Delbrouck et al., 2022). Detailed information can be found in the appendix.

4.2 Baseline and Results

We compare ConPlan with representative radiology report generation methods, including task-specific RRG models and recent medical MLLMs such as LLaVA-Med (Li et al., 2023), CheXagent (Chen et al., 2024), GPT-4v (OpenAI, 2023), Med-PaLM (Tu et al., 2024), MAIRA (Hyland et al., 2023; Bannur et al., 2024), Libra (Zhang et al., 2025)

Method	Lexical			Clinical Scores				CheXbert			
	ROUGE-L	BLEU-1	BLEU-4	RadGraph-F1	RG_{ER}	RadCliQ↓	CheXbert Sim	Micro-F1		Macro-F1	
								14	5	14	5
R2Gen	27.7	–	–	19.6	–	–	–	22.8	34.6	–	27.6
RGRG	26.4	–	–	–	–	–	–	44.7	54.7	–	–
LLaVA-Med	27.6	35.4	14.9	19.1	23.8	3.30	36.9	42.7	43.9	26.9	36.3
CheXagent	21.5	16.9	4.7	–	20.5	–	–	39.3	41.2	24.7	34.5
GPT-4V	13.2	16.4	17.8	–	13.2	–	–	35.5	25.8	20.4	19.6
Med-PaLM	27.5	32.3	11.5	26.7	–	–	–	53.6	57.9	39.8	51.6
LLaVA-Rad	30.6	38.1	15.4	–	29.4	–	–	57.3	57.4	39.5	47.7
MAIRA-1	28.9	39.2	14.2	24.3	29.6	3.10	44.0	55.7	56.0	38.6	47.7
MAIRA-2	38.4	46.5	23.4	34.6	39.6	2.64	–	58.1	59.1	41.6	50.4
Libra	36.2	51.2	24.3	32.4	36.9	2.76	46.3	55.3	58.9	40.2	52.6
FiR-Rad	32.6	–	–	32.2	–	2.89	46.8	55.1	61.3	35.5	50.2
Ours											
Qwen3-VL-2B†	39.6	43.5	22.7	33.0	35.4	2.67	50.3	56.3	59.5	42.8	52.1
+ *ConPlan*	38.3	44.1	22.4	33.1	35.3	2.65	52.0	61.7	64.2	47.4	57.6
Qwen3-VL-4B†	39.4	43.6	22.7	32.7	35.3	2.69	49.9	55.6	58.8	42.5	51.7
+ *ConPlan*	39.0	45.1	23.1	33.7	36.0	2.63	52.0	61.6	64.4	47.7	58.0
Qwen3.5-0.8B†	39.8	44.6	23.4	33.6	36.0	2.65	50.9	57.7	60.7	44.1	54.0
+ *ConPlan*	38.4	44.6	22.6	33.2	35.4	2.65	52.1	61.7	64.7	47.5	57.8
Qwen3.5-2B†	40.0	45.6	23.9	34.2	36.3	2.62	51.4	57.7	60.5	44.8	54.4
+ *ConPlan*	38.7	44.6	22.7	33.1	35.5	2.64	52.4	62.2	65.0	48.1	59.2
InternVL3.5-1B†	39.2	43.5	22.7	32.8	35.4	2.69	50.1	56.3	59.6	42.9	52.2
+ *ConPlan*	37.5	43.0	21.4	32.2	34.5	2.69	51.2	61.0	64.1	46.6	56.6
InternVL3.5-2B†	39.4	43.9	22.9	33.4	35.8	2.65	51.0	57.3	60.8	44.1	53.8
+ *ConPlan*	38.3	44.1	22.3	33.3	35.5	2.64	52.3	61.0	63.8	46.6	56.8
MiMo-VL-7B†	39.2	46.1	24.2	33.9	36.3	2.79	51.2	58.0	60.7	45.5	53.6
+ *ConPlan*	39.0	45.4	23.4	33.8	36.0	2.62	52.6	61.1	64.2	47.5	57.8
GLM-4.6V-Flash†	39.2	46.4	24.3	34.0	36.4	2.79	51.3	57.9	60.3	45.2	53.8
+ *ConPlan*	39.2	45.9	23.5	33.9	36.2	2.60	53.2	62.4	65.4	48.3	58.8

Table 1: Comparison with existing medical multimodal models on lexical and clinical report generation metrics. Underlined results denote the strongest prior result. ↓ indicates lower is better. †denotes base models directly fine-tuned with the single-turn report generation prompt template shown in Fig. 5.

and FiR-Rad (Mei et al., 2026). To evaluate plug-and-play generality, we apply ConPlan to multiple backbones, including Qwen3-VL (Bai et al., 2025), Qwen3.5 (Qwen Team, 2026), InternVL3.5 (Wang et al., 2025a), MiMo-VL (Yue et al., 2025), and GLM-4.6V-Flash (Hong et al., 2025), and compare each model with its directly fine-tuned SFT baseline.

As shown in Table 1, ConPlan brings consistent improvements in clinical correctness across most evaluated backbones, demonstrating that explicit concept planning is broadly useful beyond a single model family. Compared with the directly fine-tuned SFT baselines, ConPlan generally improves CheXbert-based F1 scores, CheXbert similarity, and clinical structure metrics, indicating better disease-label agreement and more faithful entity-relation generation. In contrast, lexical metrics such as BLEU and ROUGE show smaller or occasionally mixed changes, which is expected because ConPlan mainly optimizes clinical concept organization rather than surface-level n-gram overlap. This pattern suggests that the proposed planning mechanism primarily strengthens the clinical faithfulness of generated reports while preserving competitive textual quality.

We evaluate out-of-domain generalization on IU-Xray and CheXpert Plus in Table 2. On IU-Xray, ConPlan improves most backbones across lexical and clinical metrics, indicating that the planning mechanism transfers beyond the training distribution. On CheXpert Plus, it brings gains in clinical F1 scores for most backbones, suggesting better recognition and expression of disease-level findings under domain shift. Overall, ConPlan provides a transferable ability for clinically faithful report generation, although the improvement magnitude varies across backbones and domains.

Method	IU-Xray						CheXpert Plus					
	B-4	R-L	RG-F1	$CliQ_0$↓	^{14}Ma-F1	^{14}Mi-F1	B-4	R-L	RG-F1	$CliQ_0$↓	^{14}Ma-F1	^{14}Mi-F1
LLaVA-Rad	7.9	25.3	29.4	2.9	–	53.5	5.4	–	18.3	3.8	–	–
MAIRA-2	11.7	27.4	27.1	2.7	31.9	52.5	–	–	–	–	–	–
FiR-Rad	10.1	–	–	2.4	–	–	5.1	–	–	3.7	–	–
Ours												
Qwen3.5-0.8B†	9.4	25.3	22.2	2.7	19.0	58.7	5.8	19.9	15.2	3.7	32.3	46.9
+ *ConPlan*	11.3	26.1	22.0	2.7	23.8	59.0	6.1	20.0	15.7	3.7	36.8	52.1
Qwen3.5-2B†	9.5	24.4	19.6	2.8	22.5	59.0	6.5	20.1	15.7	3.7	34.0	48.9
+ *ConPlan*	11.3	27.0	21.9	2.7	26.4	60.5	6.0	19.9	15.7	3.6	38.0	53.2
GLM-4.6V-Flash†	8.4	23.2	19.8	3.0	20.7	56.1	6.2	19.3	15.1	3.8	34.4	48.6
+ *ConPlan*	10.0	25.0	20.9	2.8	26.8	60.2	6.6	20.3	16.0	3.6	39.6	54.6

Table 2: Performance comparison with baseline models on the IU-Xray and CheXpert Plus datasets. Best results are bolded. ↓ indicates lower is better. †denotes base models directly fine-tuned with the single-turn report generation prompt template shown in Fig. 5. The full table is in the appendix.

RSG	CCP	SCP	BLEU-4	ROUGE-L	^{14}Macro-F1	^{14}Micro-F1
✗	✗	✗	22.7	39.6	42.8	56.3
✓	✗	✗	21.5	38.8	43.2	56.9
✗	✓	✗	22.8	39.5	43.4	57.4
✓	✓	✗	21.6	38.7	43.8	58.3
✓	✗	✓	22.1	38.0	46.0	60.2
✓	✓	✓	22.4	38.3	47.4	61.7

Table 3: Ablation study of key components in ConPlan. RSG = Regional Stepwise Generation; CCP = chunk-level concept planner; SCP = step-level concept planner.

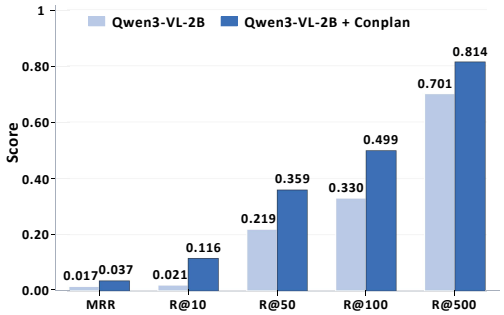


Figure 4: Clinical concept anticipation comparison between Qwen3-VL-2B and Qwen3-VL-2B + ConPlan.

4.3 Effect of Key Components

Table 3 ablates the key components of ConPlan. Compared with the direct SFT baseline, RSG alone slightly improves clinical F1 scores but lowers lexical metrics, indicating that regional supervision provides useful structure but is insufficient by itself. Adding the Chunk-level Concept Planner (CCP) improves clinical correctness, raising ^{14}Macro-F1 from 42.8 to 43.4 and ^{14}Micro-F1 from 56.3 to 57.4 while largely preserving lexical performance. The Step-level Concept Planner (SCP) brings larger clinical gains, especially with RSG, reaching 46.0 ^{14}Macro-F1 and 60.2 ^{14}Micro-F1. Using all components achieves the best clinical performance, with 47.4 ^{14}Macro-F1 and 61.7 ^{14}Micro-F1, showing that RSG, CCP, and SCP provide complementary benefits for clinically faithful generation.

4.4 Clinical Concept Anticipation

To examine whether ConPlan strengthens the anticipatory ability identified in our preliminary study, we compare Qwen3-VL-2B with Qwen3-VL-2B + ConPlan using CCAS at the start of each regional generation step. As shown in Fig. 4, ConPlan improves MRR and recall across cutoffs, indicating that upcoming positive clinical concepts are ranked higher before the corresponding findings are generated. This shows that the proposed planning objectives activate the model's latent ability to pre-organize clinically relevant concepts before surface realization.

5 Conclusion

In this paper, we study concept planning for chest X-ray report generation. We show that MLLMs exhibit latent clinical concept anticipation, but this ability is only implicitly induced by next-token prediction. To activate it, we propose ConPlan, a plug-and-play framework that reformulates report generation into a multi-step process with dual-level planning. The chunk-level planner anticipates near-future concepts in a discrete code space, while the step-level planner reuses self-generated intermediate states as priors. Experiments show that ConPlan improves report clinical correctness, and concept anticipation, highlighting the value of optimizing latent planning for faithful radiology reporting.

6 Limitations

This work has several limitations. First, although ConPlan is evaluated on multiple backbones and further tested on two out-of-domain datasets, our study is still restricted to chest X-ray report generation. We retain only frontal AP/PA radiographs and do not examine lateral views, other imaging modalities, or more heterogeneous clinical deployment scenarios. Therefore, the generality of ConPlan to broader radiology workflows, such as CT, MRI, ultrasound, or multi-view reporting, remains to be further validated.

Second, Regional Stepwise Generation relies on automatic report decomposition to construct region-level supervision. Although the parser is instructed to preserve modifiers, temporal comparisons, and diagnostic reasoning, imperfect decomposition may introduce noise into the structured training targets. The fixed regional order also provides a simplified organization of radiology reports and may not fully match all reporting styles or institution-specific templates. Future work can explore more adaptive region selection and decomposition strategies.

Third, ConPlan is a plug-and-play framework for autoregressive medical MLLMs, but it still requires access to intermediate decoder hidden states. As a result, it is mainly applicable to open-source or internally accessible models, rather than black-box API-only systems. The multi-step generation protocol and auxiliary planning module also introduce additional training and inference overhead compared with standard single-pass report generation.

7 Ethical Considerations

This work studies automatic chest X-ray report generation with medical MLLMs. The proposed method is intended to support research on clinically faithful report generation, rather than to replace radiologists or serve as an autonomous diagnostic system. Although ConPlan improves clinical correctness under automatic evaluation, generated reports may still contain omissions, hallucinated findings, incorrect negations, or misleading temporal comparisons. Therefore, any practical use of such systems should require expert review, clear human oversight, and appropriate uncertainty communication.

All experiments are conducted on public research datasets with de-identified radiology images and reports. We do not attempt to recover patient identity or use protected health information beyond the scope permitted by the datasets. When releasing code, processed annotations, or model outputs, care should be taken to follow the original dataset licenses and avoid exposing potentially sensitive clinical text. In addition, models trained on historical radiology data may inherit dataset biases related to patient demographics, acquisition protocols, reporting styles, and institutional practice patterns. These biases could affect performance across populations or clinical environments. Future deployment-oriented studies should include subgroup analysis, external validation, radiologist evaluation, and prospective safety assessment before considering real-world clinical integration.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, and 1 others. 2024. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P Langlotz. 2024. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. *arXiv preprint arXiv:2405.19538*.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 5904–5914.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. In *EMNLP*, pages 1439–1449.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, and 1 others. 2024. Chexagent: Towards a foundation model for chest x-ray interpretation. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*.

Yashin Dicente Cid, Matthew Macpherson, Louise Gervais-Andre, Yuanyi Zhu, Giuseppe Franco, Ruggiero Santeramo, Chee Lim, Ian Selby, Keerthini Muthuswamy, Ashik Amlani, and 1 others. 2024. Development and validation of open-source deep neural networks for comprehensive chest X-ray reading: a retrospective, multicentre study. *The Lancet Digital Health*, 6(1):e44–e57.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. 2022. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360.

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

Tiancheng Gu, Kaicheng Yang, Xiang An, Ziyong Feng, Dongnan Liu, and Weidong Cai. 2025. ORID: Organ-regional information driven framework for radiology report generation. In *WACV*, pages 378–387.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, and 1 others. 2025. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*.

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

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, and 1 others. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, and 1 others. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*.

Haibo Jin, Haoxuan Che, Sunan He, and Hao Chen. 2025. A chain of diagnosis framework for accurate and explainable radiology report generation. *IEEE Transactions on Medical Imaging*.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. *NeurIPS*, 36:28541–28564.

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

Chang Liu, Yuanhe Tian, Weidong Chen, Yan Song, and Yongdong Zhang. 2024. Bootstrapping large language models for radiology report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18635–18643.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2021. Exploring and distilling posterior and prior knowledge for radiology report generation. In *CVPR*, pages 13753–13762.

Yihe Liu, Huibin Wang, Xianming Hu, Pinyi Zhang, Jiahao Xiong, Chenglin Wang, Nuoyi Chen, Hongbo Zhao, Jie Zhang, and Kai Zhang. 2026a. Next-tobe: Probabilistic next token-bag exploitation for activating anticipatory capacity in llms. In *The Fourteenth International Conference on Learning Representations*.

Yuliang Liu, Yunchong Song, Yixuan Wang, Kewen Ge, Alex Lamb, Qipeng Guo, Kai Chen, Bowen Zhou, and Zhouhan Lin. 2026b. Next concept prediction in discrete latent space leads to stronger language models. *arXiv preprint arXiv:2602.08984*.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge university press.

Xin Mei, Libin Yang, Dehong Gao, Xiaoyan Cai, Junwei Han, and Tianming Liu. 2026. Fir-rad: Fine-grained reinforcement with structured reasoning for chest x-ray report generation. *IEEE Transactions on Medical Imaging*, pages 1–1.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine learning for health (ML4H)*, pages 353–367. PMLR.

OpenAI. 2023. GPT-4V(ision) system card.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Chantal Pellegrini, Ege Özsoy, Benjamin Busam, Nassir Navab, and Matthias Keicher. 2023. Radialog: A large vision-language model for radiology report generation and conversational assistance. *arXiv preprint arXiv:2311.18681*.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Pranav Rajpurkar, Jeremy Irvin, Robyn L Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P Langlotz, and 1 others. 2018. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Medicine*, 15(11):e1002686.

Jarrel CY Seah, Cyril HM Tang, Quinlan D Buchlak, Xavier G Holt, Jeffrey B Wardman, Anuar Aimoldin, Nazanin Esmaili, Hassan Ahmad, Hung Pham, John F Lambert, and 1 others. 2021. Effect of a comprehensive deep-learning model on the accuracy of chest X-ray interpretation by radiologists: a retrospective, multireader multicase study. *The Lancet Digital Health*, 3(8):e496–e506.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew Lungren. 2020. 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)*, pages 1500–1519.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. 2023. Interactive and explainable region-guided radiology report generation. In *CVPR*, pages 7433–7442.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, and 1 others. 2024. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138.

Ellen M Voorhees. 1999. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025a. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Xinyi Wang, Grazziela Figueredo, Ruizhe Li, Wei Emma Zhang, Weitong Chen, and Xin Chen. 2025b. A survey of deep-learning-based radiology report generation using multimodal inputs. *Medical Image Analysis*, page 103627.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023a. METransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *CVPR*, pages 11558–11567.

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

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

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y Ng, and 1 others. 2023. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9).

Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, and 1 others. 2025. Mimo-vl technical report. *arXiv preprint arXiv:2506.03569*, 5.

Xi Zhang, Zaiqiao Meng, Jake Lever, and Edmond SL Ho. 2025. Libra: Leveraging temporal images for biomedical radiology analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17275–17303.

A Experimental details

A.1 Datasets

We conduct experiments on the MIMIC-CXR dataset (Johnson et al., 2019), a large-scale public chest X-ray dataset containing radiographic images and corresponding free-text radiology reports. Following common practice in chest X-ray report generation, we retain only frontal chest radiographs, including AP (anteroposterior) and PA (posteroanterior) views, and exclude other views such as lateral images to reduce view-dependent variation and construct a more consistent generation setting. For each sample, the model takes the current chest X-ray image and associated clinical context as input. When a prior study is available, the prior image and prior1 report are also provided to support longitudinal comparison. The supervision target is the findings section of the current radiology report. In our structured multi-step setting, each original image-to-report sample is further converted into a multi-turn instruction-following format, where the model sequentially generates section-level findings for lungs, heart/mediastinum, pleura, bones/soft tissues, and devices/lines and tubes, and then integrates these intermediate outputs into a complete report. The target of the final step remains the original reference report, ensuring that the structured

supervision is aligned with the standard report generation objective.

A.2 Evaluation Metrics

We evaluate model performance from both text generation quality and clinical correctness perspectives. For textual similarity, we report BLEU-1/4 (Papineni et al., 2002), and ROUGE-L (Lin, 2004), which measure unigram matching, longer n-gram consistency, and longest-common-subsequence-based sequence similarity, respectively. Since these metrics mainly capture surface-level textual overlap, we further adopt several clinically oriented metrics. Specifically, we report **CheXbert** vector similarity (Smit et al., 2020; Yu et al., 2023), which measures report-level clinical semantic similarity using representations from the CheXbert model. We also report **F1-CheXbert**, i.e., F1 over CheXpert-style labels (Irvin et al., 2019) extracted by the CheXbert labeler (Smit et al., 2020). This includes micro-F1 and macro-F1 over 14 thoracic observations **MicroF1-14** and **MacroF1-14** and over 5 key observations **MicroF1-5** and **MacroF1-5**. In addition, we use **RadGraph-F1** (Jain et al., 2021) to evaluate the structural overlap of clinical entities and relations, $\mathbf{RadCliQ}_0$ (Yu et al., 2023) to estimate report-level clinical errors, and $\mathbf{RG}_{\mathbf{ER}}$ (Delbrouck et al., 2022) as a complementary factual correctness metric for radiology report generation. Together, these metrics provide a comprehensive assessment of surface-level similarity, clinical semantic consistency, disease-label agreement, entity-relation correctness, and overall clinical quality.

A.3 Data Template

Fig. 5 shows the prompt template used for single-turn radiology report generation. The model receives a paired prior and current chest X-ray image, together with structured clinical context including indication, history, comparison, technique, and the prior report. Based on this information, the model is instructed to generate a complete radiology report in one step while explicitly comparing the current image with the prior image. This setting evaluates the model's ability to integrate visual evidence, clinical context, and temporal changes without intermediate section-wise reasoning.

Single-turn Report Generation Prompt Template

Input
```
Prior Image: <image>
Current Image: <image>
```

Context
```
Indication: {indication}
History: {history}
Comparison: {comparison}
Technique: {technique}
Prior Report: {prior_report}
```

Instruction
Generate a comprehensive radiology report. Compare the current image with the prior image, noting any significant changes, progression, or stability.

Assistant
```
{final_report}
```

Figure 5: Prompt template for single-turn radiology report generation. The model directly generates the complete target report from the image pair and clinical context.

Fig. 6 illustrates the structured multi-step prompt template used for radiology report generation. Given the paired prior and current images and the same clinical context as the single-turn setting, the model first generates section-specific findings for lungs, heart/mediastinum, pleura, bones/soft tissues, and devices/lines/tubes. The final step then integrates these intermediate findings into a complete and coherent report while preserving relevant comparison statements. This design encourages explicit anatomical decomposition and temporal reasoning before final report synthesis.

A.4 Experimental Parameters

Table 4 summarizes the main experimental parameters used in our local NCP setting. We fine-tune MLLMs with LoRA while keeping the vision tower trainable. The proposed NCP module is inserted at the 4-th language-model layer and operates on 6-token chunks. Both global and local NCP are enabled, with token-shift and decode-time injection used during inference. The final objective combines next-token prediction, NCP prediction, vector-quantization, and local NCP losses with equal auxiliary weights of 0.1. All models are trained for 3 epochs using bf16 precision, FlashAttention-2, and DeepSpeed ZeRO-2 on four H20 GPUs.

Structured Multi-step Prompt Template

Input

```
Prior Image: <image>
Current Image: <image>
```

Context

```
Indication: {indication}
History: {history}
Comparison: {comparison}
Technique: {technique}
Prior Report: {prior_report}
```

Instruction

Generate a comprehensive radiology report. Compare the current image with the prior image, noting any significant changes, progression, or stability.

Step 1/6: LUNGS
Please generate ONLY the LUNGS findings based on the images and clinical context above. If there is no relevant content, output <EMPTY>.
Assistant: {lungs_findings}

Step 2/6: HEART/MEDIASTINUM
Please generate ONLY the HEART/MEDIASTINUM findings. If none, output <EMPTY>.
Assistant: {heart_mediastinum_findings}

Step 3/6: PLEURA
Please generate ONLY the PLEURA findings. If none, output <EMPTY>.
Assistant: {pleura_findings}

Step 4/6: BONES/SOFT TISSUES
Please generate ONLY the BONES/SOFT TISSUES findings. If none, output <EMPTY>.
Assistant: {bones_soft_tissues_findings}

Step 5/6: DEVICES/LINES&TUBES
Please generate ONLY the DEVICES/LINES&TUBES findings (if any). If none, output <EMPTY>.
Assistant: {devices_findings}

Step 6/6: Final Report
Now write a complete, coherent radiology report by integrating all the sections above (keep comparison statements if present).
Assistant: {final_report}

Figure 6: Prompt template for structured multi-step radiology report generation. The first five steps generate section-level findings, and the final step synthesizes them into a complete radiology report.

Table 4: Experimental parameters for the local NCP setting.

Parameter	Value
Backbone model	Qwen3-VL-2B-Instruct
Training stage	Supervised fine-tuning
Fine-tuning method	LoRA
LoRA rank / alpha	64 / 128
LoRA target modules	All target modules
Additional saved modules	`visual.merger`, `ncp_module`
Vision tower	Trainable
Template	`qwen3_vl_nothink`
Maximum image pixels	147,456
Maximum sequence length	2,048
Planner type	Global Planner + local Planner
Planner insertion layer	4
Planner chunk size	6 tokens
Planner transformer layers	2
Planner codebook size	64
Planner token shift	Enabled
Decode-time Planner injection	Enabled
Local Planner inference	Enabled
Local Planner prior weight	1.0
Planner loss scope	Assistant tokens
Planner loss weight	0.1
Planner loss weight	0.1
Local Planner loss weight	0.1
Planner commitment coefficient	0.25
Precision	bf16
Attention implementation	FlashAttention-2
Distributed training	DeepSpeed ZeRO-2
Number of GPUs	4
Per-device batch size	4
Gradient accumulation steps	2
Effective batch size	32
Learning rate	1.0×10^{-4}
LR scheduler	Cosine
Warmup ratio	0.03
Weight decay	0
Training epochs	3
Total training steps	15,279
Logging interval	10 steps
Save interval	500 steps
Evaluation interval	500 steps
Best checkpoint	Step 15,000
Inference checkpoint	Step 15,279
Maximum new tokens	512
KV cache at inference	Enabled

A.5 Full out-of-domain validation Experiments

Table 5 presents the full out-of-domain validation results on IU-Xray and CheXpert Plus. We evaluate conplan across multiple backbone families, including Qwen3-VL, Qwen3.5, InternVL3.5, MiMo-VL, and GLM-4.6V-Flash, and compare each model with its single-turn fine-tuned counterpart. Overall, conplan consistently improves most metrics across both datasets, indicating that the proposed structured generation strategy generalizes beyond the training domain.

On IU-Xray, conplan improves BLEU-4 and ROUGE-L for all listed backbones, showing better lexical quality and report-level overlap. It also reduces CliQ_0 in most settings, suggesting fewer clinical quality errors. On CheXpert Plus, the improvements are most evident in clinical F1 scores: conplan substantially increases both macro-F1 and micro-F1 for nearly all backbones. These gains demonstrate that conplan improves not only textual similarity but also clinically relevant finding prediction under out-of-domain evaluation.

Method	IU-Xray						CheXpert Plus					
	B-4	R-L	RG-F1	CliQ$_0$↓	^{14}Ma-F1	^{14}Mi-F1	B-4	R-L	RG-F1	CliQ$_0$↓	^{14}Ma-F1	^{14}Mi-F1
LLaVA-Rad	7.9	25.3	29.4	2.9	–	53.5	5.4	–	18.3	3.8	–	–
MAIRA-2	11.7	27.4	27.1	2.7	31.9	52.5	–	–	–	–	–	–
FiR-Rad	10.1	–	–	2.4	–	–	5.1	–	–	3.7	–	–
Ours												
Qwen3-VL-2B†	9.6	24.9	20.9	2.8	21.3	57.4	6.2	19.9	15.0	3.7	31.4	46.6
+ *ConPlan*	11.5	26.6	22.2	2.7	24.3	57.4	6.0	19.9	16.0	3.6	37.2	53.0
Qwen3-VL-4B†	8.9	24.8	20.1	2.8	16.6	55.3	5.6	19.1	13.9	3.8	30.2	43.3
+ *ConPlan*	11.3	25.5	21.5	2.7	25.9	59.2	6.4	20.2	15.9	3.6	37.4	52.9
Qwen3.5-0.8B†	9.4	25.3	22.2	2.7	19.0	58.7	5.8	19.9	15.2	3.7	32.3	46.9
+ *ConPlan*	11.3	26.1	22.0	2.7	23.8	59.0	6.1	20.0	15.7	3.7	36.8	52.1
Qwen3.5-2B†	9.5	24.4	19.6	2.8	22.5	59.0	6.5	20.1	15.7	3.7	34.0	48.9
+ *ConPlan*	11.3	27.0	21.9	2.7	26.4	60.5	6.0	19.9	15.7	3.6	38.0	53.2
InternVL3.5-1B†	10.7	25.4	21.2	2.7	17.9	57.7	6.1	19.3	14.2	3.8	30.0	44.5
+ *ConPlan*	11.4	26.7	22.0	2.7	17.8	59.0	6.2	20.1	16.1	3.7	36.6	52.2
InternVL3.5-2B†	9.2	24.0	18.7	2.9	20.7	58.1	6.2	20.2	15.4	3.7	32.8	47.9
+ *ConPlan*	10.9	25.6	20.9	2.7	23.1	60.6	6.1	20.2	15.8	3.6	38.4	53.6
MiMo-VL-7B†	8.2	23.2	19.5	3.0	18.1	55.8	5.9	19.1	14.6	3.9	32.7	47.2
+ *ConPlan*	10.8	25.1	20.1	2.8	21.2	57.8	5.9	19.8	15.3	3.7	38.1	52.9
GLM-4.6V-Flash†	8.4	23.2	19.8	3.0	20.7	56.1	6.2	19.3	15.1	3.8	34.4	48.6
+ *ConPlan*	10.0	25.0	20.9	2.8	26.8	60.2	6.6	20.3	16.0	3.6	39.6	54.6

Table 5: Performance comparison with baseline models on the IU-Xray and CheXpert Plus datasets. Best results are bolded. ↓ indicates lower is better. †denotes base models directly fine-tuned with the single-turn report generation prompt template shown in Fig. 5.

B Regional Stepwise Generation

Directly generating a complete radiology report in a single pass provides only coarse sequence-level supervision and leaves the organization of regional findings implicit. To expose clinically meaningful generation steps, we introduce *Regional Stepwise Generation* (RSG), a structured protocol that decomposes chest X-ray reporting into region-level finding generation followed by report-level synthesis.

As shown in Fig. 7, RSG starts from the original sample, which contains chest X-ray images, clinical context, prior report when available, and the target findings section. We use Qwen3-VL-8B-Instruct as an automatic report parser to decompose the target report into five anatomical or functional regions: lungs, heart/mediastinum, pleura, bones/soft tissues, and devices/lines and tubes. The parser is instructed to preserve modifiers, temporal comparisons, and diagnostic reasoning, while normalizing unmentioned regions to <EMPTY>.

Based on the decomposed sections, each original single-turn sample is converted into a six-step instruction-following sequence. In Steps 1–5, the model generates one region-specific finding at a time following a fixed regional order. In Step 6, it integrates the generated regional findings into a complete and coherent report. The supervision target of the final step remains the original findings section, ensuring that RSG remains aligned with the standard report-generation objective.

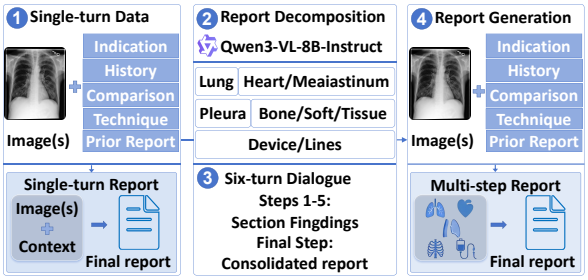


Figure 7: Overview of *Regional Stepwise Generation* (RSG). An original single-turn image-to-report sample is first decomposed into region-level findings and then reformulated as a six-step dialogue. The model generates findings for five anatomical or functional sections and finally synthesizes them into a complete report.

RSG is used consistently during both training and inference. During training, it provides explicit region-level supervision and defines the stepwise state-to-finding structure used by the step-

level planner. During inference, the model follows the same stepwise protocol, first producing region-specific findings and then synthesizing the final report. Therefore, RSG is not merely a data preprocessing strategy; it serves as the structured generation backbone on which the chunk-level and step-level concept planners operate.

C Related Work

Early RRG systems typically incorporated clinical semantics into standard encoder-decoder architectures, improving report completeness and clinical accuracy through memory mechanisms, medical knowledge modeling, cross-modal alignment, multiple expert tokens, and region-level visual grounding (Chen et al., 2020, 2021; Liu et al., 2021; Wang et al., 2023a; Tanida et al., 2023). The development of foundation models has driven a paradigm shift in RRG from task-specific encoder-decoder pipelines to generalist medical Multimodal Large Language Model (MLLM) frameworks. For instance, R2GenGPT (Wang et al., 2023b) connects visual encoders to frozen LLMs via efficient vision-language alignment to reduce training costs, while Bootstrapping LLMs for RRG (Liu et al., 2024) further explores how to leverage large language models for cross-modal report generation under limited radiology data. Meanwhile, radiology-specific MLLMs such as RaDialog (Pellegrini et al., 2023) and MAIRA (Hyland et al., 2023) adapt large vision-language models to chest X-ray report generation and interactive diagnostic assistance. More broadly, general medical MLLMs such as LLaVA-Med (Li et al., 2023), Med-Flamingo (Moor et al., 2023), and RadFM (Wu et al., 2025) have demonstrated strong instruction-following, few-shot generalization, and multimodal medical understanding capabilities, providing stronger language modeling and clinical knowledge representation foundations for modern radiology report generation.

Despite these advances, existing medical MLLMs still largely rely on the standard autoregressive next-token prediction objective, which mainly encourages local token-level fitting. However, radiology report generation requires the model to identify upcoming clinical concepts, plan descriptions across different anatomical or functional regions, and maintain consistency between local findings and the overall report. As a result, even models with implicit reasoning ability may still suffer from concept omission, disorganized regional descriptions, or inconsistent statements in long-form report generation. How to explicitly model and leverage the latent clinical concept planning ability of MLLMs, so that they can organize future clinical semantics during autoregressive generation, remains a key challenge for faithful radiology report generation.

D Sensitivity to Hyperparameters

Table 6 reports the sensitivity of ConPlan to the inserted decoder layer and chunk size. For the insertion layer, L4 achieves the best overall performance, with the highest ROUGE-L, ^{14}Macro-F1, and ^{14}Micro-F1. In comparison, inserting the planner at L1 gives slightly lower clinical scores, while L8 leads to a clearer drop in both Macro-F1 and Micro-F1. This suggests that middle decoder layers provide a better trade-off between local token representation and higher-level semantic organization, making them more suitable for concept planning.

For chunk size, ConPlan remains relatively stable across $C = 4$, $C = 6$, and $C = 8$. The setting $C = 6$ obtains the best ^{14}Macro-F1 and matches the best ^{14}Micro-F1, while $C = 4$ gives nearly identical lexical performance. A larger chunk size of $C = 8$ slightly improves ROUGE-L but weakens clinical F1 scores, indicating that overly coarse chunks may blur fine-grained clinical concept transitions. Based on these results, we use L4 and $C = 6$ as the default configuration.

Study	Setting	BLEU-4	ROUGE-L	^{14}Macro-F1	^{14}Micro-F1
Layer	L1	22.2	38.1	46.0	61.6
	L4	22.4	38.3	47.4	61.7
	L8	22.4	37.8	45.3	60.9
Chunk Size	$C = 4$	22.4	38.3	46.9	61.7
	$C = 6$	22.4	38.3	47.4	61.7
	$C = 8$	22.3	38.4	46.5	61.1

Table 6: Sensitivity analyses on the inserted decoder layer and chunk size. Layer denotes where the concept planning module is inserted, and C denotes the number of tokens grouped into each chunk. F1 scores are reported as percentages.

E Why Structural Metrics Change Modestly

The improvements on RadGraph-F1 and $\mathrm{RG}_{\mathrm{ER}}$ are less pronounced than those on CheXbert-based metrics. This is mainly because ConPlan is designed to improve clinical concept organization and disease-level correctness, rather than directly optimizing entity-relation tuples or surface-level

factual error patterns. RadGraph-F1 relies on exact or near-exact matching of extracted clinical entities and relations, making it sensitive to wording, report style, and regional rephrasing introduced by the multi-step generation protocol. Similarly, $\mathrm{RG}_{\mathrm{ER}}$ reflects report-level factual errors derived from structured extraction, and may not fully capture improvements in concept anticipation or disease-label consistency when the generated report expresses the same finding with different phrasing. Therefore, the relatively stable RadGraph-F1 and $\mathrm{RG}_{\mathrm{ER}}$ scores suggest that ConPlan primarily strengthens disease-level clinical faithfulness, while structural entity-relation alignment remains a complementary direction for future optimization.

F Future Work

Future work will extend ConPlan from disease-level concept planning to more structure-aware clinical reasoning. In particular, incorporating explicit entity-relation planning or RadGraph-style supervision may help improve structural metrics such as RadGraph-F1 and $\mathrm{RG}_{\mathrm{ER}}$, which are not directly optimized in the current framework. Another direction is to make the regional generation protocol more adaptive, allowing the model to select clinically relevant regions dynamically instead of following a fixed anatomical order. Beyond chest X-rays, validating ConPlan on other imaging modalities and multi-view or longitudinal reporting scenarios would further test whether concept planning provides a general mechanism for faithful medical report generation.

G Dataset Statistics

Table 7 summarizes the statistics of the processed datasets used in our experiments. The in-domain training, development, and test sets are constructed from MIMIC-CXR, while the out-of-domain test sets are derived from the official test splits of CheXpert Plus and IU-Xray annotations. Specifically, CheXpert Plus uses the test split from `chexpert_annotation.json`, and IU-Xray uses the test split from `iu_annotation.json`. All processed samples are converted into the structured six-step generation format, resulting in 12 dialogue messages per example.

For MIMIC-CXR, the processed in-domain data contain 166,708 examples in total, including 162,959 training examples, 1,286 development examples, and 2,463 test examples.

Dataset	Split	Examples	Current-only	Prior-current	Msgs.	Avg./Med. Words
MIMIC-CXR	Train	162,959	67,738	95,221	12	49.4 / 46
MIMIC-CXR	Dev	1,286	513	773	12	49.1 / 46
MIMIC-CXR	Test	2,463	344	2,119	12	61.5 / 57
IU-Xray	Test	590	590	0	12	28.3 / 25
CheXpert Plus	Test	11,562	7,167	4,395	12	63.9 / 55

Table 7: Statistics of the processed datasets used in our experiments. "Current-only" denotes samples containing only the current image, while "Prior-current" denotes temporal samples containing both prior and current images. "Msgs." is the number of dialogue messages per example. Avg./Med. Words reports the average and median word counts of the final target report.

For out-of-domain validation, we use 590 IU-Xray test examples from `iu_annotation.json` and 11,562 CheXpert Plus test examples from `chexpert_annotation.json`. The processed files use the `noEMPTY_noNegative` setting to remove empty and negative training targets.

H Example of Reducing Pneumonia Hallucination

Figure 8 shows a case where the single-turn baseline hallucinates pneumonia. The reference report indicates clear lungs with no pneumothorax or pleural effusion, while the no-NCP baseline incorrectly predicts a new left lower-lobe opacity suspicious for pneumonia. In contrast, our model correctly reports clear lungs and preserves the key negative findings, reducing this clinically relevant false positive.

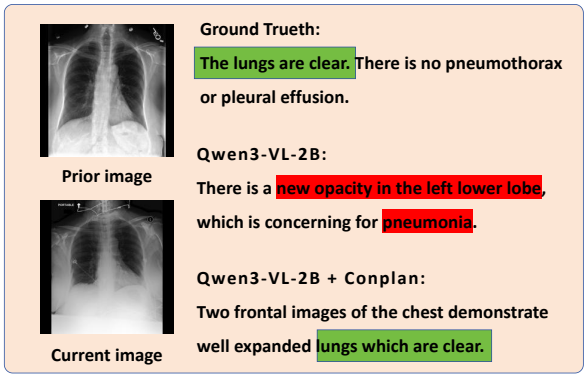


Figure 8: Example of reducing pneumonia hallucination. The no-NCP baseline falsely reports a new left lower-lobe opacity suspicious for pneumonia, whereas our model matches the reference by describing clear lungs without pneumothorax or pleural effusion.