GDPO-based Multi-Alignment Reward Optimization for Clinically Grounded Medical Image Captioning

ImageCLEF Lab at CLEF 2026

Hyun Jun Kim[1,†], Heeseung Shin[2,†], Minjun Kim[3] and Changwon Lim[3,*]

[1]Department of Smart Cities, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Republic of Korea
[2]Department of AI in Medicine, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Republic of Korea
[3]Department of Statistics and Data Science, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Republic of Korea

Abstract

Medical image captioning requires models to generate descriptions that are linguistically fluent, visually grounded, and clinically faithful. This paper describes the AI Stat Lab system for the CLEF 2026 ImageCLEFmedical Caption task, submitted under the official leaderboard username `AIstatLab`. The system combines supervised fine-tuning with post-SFT Group Reward-Decoupled Policy Optimization (GDPO). To move beyond token-level imitation, we design a multi-alignment reward organized around three complementary axes and implemented with four reward components: BERTScore Recall and ROUGE-1 for caption-to-caption alignment, BiomedCLIP-based similarity for caption-to-image alignment, and a frozen biomedical concept extractor for caption-to-CUI alignment. We additionally use reward scheduling to balance heterogeneous reward signals, applying a piecewise schedule for the standard track and a dynamic schedule for the synthetical track. Final submissions were generated from selected checkpoints using deterministic decoding and a conservative rule-based fallback selector for malformed outputs. Our submissions ranked first by overall score on the official leaderboards for both the standard and synthetical caption prediction tracks. These results suggest that medical-domain initialization, multi-objective post-SFT optimization, and carefully controlled inference-time selection provide an effective framework for competitive medical image captioning. The source code and implementation details are publicly available at: https://github.com/HJunK0410/AIStatLab-CLEF2026_MIC.

Keywords

Medical image captioning, Vision-language models, Reinforcement learning, Policy optimization, Multi-objective reward, Reward scheduling, Clinical concept alignment

1. Introduction

Medical image captioning is a multimodal generation task that requires both visual understanding and clinically faithful language generation. Unlike general image captioning, which often focuses on salient objects, actions, and scene-level relations in open-domain images, medical image captioning must capture clinically relevant information such as imaging modality, anatomical structures, abnormal findings, disease-related patterns, and visually supported observations [1]. A generated caption may be fluent and plausible while remaining clinically unreliable if it omits important findings, introduces unsupported concepts, or describes image findings that are not visually grounded. Therefore, medical image captioning systems should be optimized not only for linguistic fluency, but also for visual grounding, clinical factuality, and concept-level completeness.

ImageCLEFmedical Caption 2026 was organized as part of ImageCLEF, a CLEF benchmarking campaign for multimedia retrieval tasks including medical, social media, and content recommendation applications [2]. The medical captioning task provides a shared evaluation setting with standard Caption Prediction and Caption Prediction Synthetical tracks [3]. Recent ImageCLEFmedical captioning tasks

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
†Hyun Jun Kim and Heeseung Shin contributed equally to this work.
hyunjun0615@cau.ac.kr (H. J. Kim); hs970416@cau.ac.kr (H. Shin); goodwill1669@cau.ac.kr (M. Kim); clim@cau.ac.kr (C. Lim)
0009-0000-7512-7987 (H. J. Kim); 0009-0004-0430-9547 (H. Shin); 0009-0008-0659-0941 (M. Kim); 0000-0002-3182-8420 (C. Lim)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

are closely connected to large-scale medical image-caption resources such as ROCOv2 [4] and evaluate generated captions using multiple metrics that capture complementary aspects of caption quality. This multi-dimensional evaluation setting motivates captioning systems that consider reference-caption similarity, lexical consistency, visual grounding, and clinical concept preservation rather than optimizing only a single text-similarity objective.

A common approach to image-conditioned text generation is supervised fine-tuning (SFT) with teacher forcing, where the model learns to predict reference-caption tokens conditioned on the input image and prompt. Although this objective is stable and effective for learning medical caption style, modality-specific terminology, and common anatomical descriptions, token-level likelihood does not directly optimize sequence-level caption quality. As a result, an SFT-only model may generate fluent captions that are incomplete, weakly grounded, or insufficiently faithful to clinically important concepts.

Sequence-level optimization offers a natural way to reduce this mismatch between token-level training and sequence-level evaluation. Earlier image captioning work, such as self-critical sequence training, showed that reinforcement learning can optimize captioning models using task-level rewards computed on complete generated captions [5]. More recent LLM post-training methods compare multiple generated outputs to form relative advantages and improve generation behavior beyond maximum-likelihood imitation [6, 7]. These methods provide a useful framework for refining SFT captioning models using rewards that reflect semantic similarity, lexical overlap, visual grounding, and clinical concept consistency.

In this work, we adopt a post-SFT Group Reward-Decoupled Policy Optimization (GDPO) pipeline for medical image captioning. Our central design choice is a multi-alignment reward organized around three complementary axes and implemented using four reward components. Caption-to-caption alignment is represented by BERTScore Recall [8] and ROUGE-1 [9], which encourage semantic content recovery and lexical consistency with reference captions. Caption-to-image alignment is represented by BiomedCLIP-based image-text similarity [10], which serves as a training-time proxy for visual grounding. Caption-to-CUI alignment is represented by a frozen biomedical concept extractor, which encourages coverage of clinically relevant concepts using UMLS-based concept representations and biomedical language modeling resources [11, 12, 13, 14].

Combining these rewards introduces an additional optimization challenge because the reward components may differ in scale, noise level, sparsity, and reliability. Moreover, the relative usefulness of each reward may change across training stages and task settings. Fixed reward weights are simple and reproducible, but they may not adequately handle reward conflict or data-regime differences. We therefore investigate reward scheduling as a practical mechanism for balancing heterogeneous reward signals during GDPO. Specifically, we use a manually designed piecewise schedule for the standard task setting and an online dynamic schedule for the synthetical task setting.

The main contributions of this work are as follows:

- We develop a two-stage medical image captioning system that combines teacher-forcing SFT with post-SFT GDPO optimization.
- We design a multi-alignment reward organized around caption-to-caption, caption-to-image, and caption-to-CUI alignment, implemented through BERTScore Recall, ROUGE-1, BiomedCLIP similarity, and a frozen biomedical CUI extractor.
- We investigate reward scheduling strategies for balancing heterogeneous reward components across the standard and synthetical caption prediction settings.
- We report first-place systems by overall score on the official ImageCLEFmedical Caption 2026 leaderboards for both task settings, while analyzing the effects of reward design, decoding, and conservative rule-based post-processing at the system level.

2. Related Work

2.1. Medical Image Captioning and ImageCLEFmedical Caption

Image captioning was initially developed in the general vision-language setting, where encoder–decoder models generate natural-language descriptions from visual inputs. Representative early systems include Show and Tell, which formulates caption generation as image-conditioned sequence modeling, and Show, Attend and Tell, which incorporates visual attention to improve region-level grounding during generation [15, 16]. Medical image captioning extends this problem from open-domain scene description to clinically grounded image interpretation. In this setting, a caption should not only be grammatically well formed, but also clinically meaningful, visually supported, and precise with respect to imaging modality, anatomy, abnormal findings, and disease-related observations [1]. This makes medical image captioning particularly challenging because a generated caption may appear fluent while still omitting important findings or introducing unsupported clinical concepts.

The ImageCLEFmedical Caption series has provided a standardized shared-task setting for medical image caption prediction and medical concept detection. Recent editions are closely connected to large-scale medical image–caption resources such as ROCOv2, which provides multimodal radiology data with captions and associated biomedical concepts [4]. The 2026 edition further defines standard and synthetical caption prediction settings and evaluates systems under shared data, task definitions, and official scoring protocols [3, 17]. This benchmark setting is important because medical captioning systems must be assessed not only by general language quality, but also by their ability to preserve clinically relevant visual and textual information.

Within the ImageCLEFmedical Caption benchmark, prior systems have explored a wide range of design choices, including model backbone selection, domain-specific pretraining, supervised fine-tuning, prompt design, decoding strategies, and inference-time refinement [18, 19]. These approaches show that medical captioning performance depends on both the underlying vision-language model and the procedure used to select or refine generated captions. However, many systems primarily rely on supervised fine-tuning and inference-time selection, which can improve the choice among generated candidates but does not directly update the model distribution according to sequence-level evaluation criteria.

Our work follows this shared-task setting while focusing on post-SFT optimization for medical image captioning. Instead of using metric-aware signals only for inference-time selection, we use them as training-time reward signals within a GDPO framework. The foundation models, sequence-level optimization methods, and evaluation-oriented reward signals that motivate this design are reviewed in the following subsections.

2.2. Vision-Language Foundation Models for Medical Imaging

Vision-language foundation models have substantially improved multimodal representation learning and image-conditioned generation. CLIP introduced large-scale contrastive image–text pretraining and showed that visual representations can be learned from natural-language supervision at scale [20]. In the biomedical domain, BiomedCLIP adapts this contrastive learning paradigm to biomedical images and scientific text, providing domain-specific image–text representations that are better aligned with biomedical visual concepts than general-domain models [10]. Such models are particularly useful for medical image captioning because the task requires not only visual recognition, but also alignment between image evidence and clinically meaningful language.

Recent vision-language models have further extended this paradigm from representation learning to open-ended image-conditioned generation. The Qwen-VL model family has shown strong multimodal perception and generation capabilities, with later variants improving visual resolution handling, document and chart understanding, and general-purpose image-conditioned instruction following [21, 22, 23]. In parallel, medical-domain vision-language models such as MedGemma have been developed to support medical image and text understanding, reflecting the growing importance of domain-adapted multimodal models for clinical and biomedical applications [24]. Open medical VLM

checkpoints, including Qwen-based MedVL models, provide another practical route for adapting strong general-purpose multimodal backbones to medical image interpretation tasks [25].

In our system, vision-language foundation models are used in two complementary ways. First, Qwen-based and MedGemma-based models serve as caption generation backbones that are supervised fine-tuned and subsequently optimized with GDPO. Second, BiomedCLIP is used as a reward-side representation model to compute an image–text similarity signal during post-SFT optimization. This dual use reflects the practical role of medical VLMs both as generative captioning models and as auxiliary evaluators of visual grounding.

However, strong multimodal initialization alone does not determine how a captioning model should balance semantic faithfulness, lexical consistency, visual grounding, and clinical concept coverage. A model may generate fluent and medically plausible text while still omitting important findings or introducing unsupported clinical details. This motivates post-SFT optimization on top of medical VLM backbones, where sequence-level reward signals can be used to refine the model toward the multi-dimensional evaluation criteria of medical image captioning.

2.3. Sequence-Level Optimization for Captioning and Generation

Teacher forcing is the standard training objective for image-conditioned text generation. It optimizes token-level likelihood by increasing the probability of reference tokens given an image and textual prompt. This objective is stable and effective for supervised fine-tuning, but it does not directly optimize the sequence-level metrics used to evaluate generated captions. In medical image captioning, this mismatch is particularly important because many clinically relevant errors, such as omitted findings, incorrect anatomical descriptions, unsupported disease mentions, or weak visual grounding, are properties of the complete generated caption rather than isolated token predictions.

Sequence-level optimization provides a way to reduce this mismatch by assigning rewards to complete generated outputs. In image captioning, self-critical sequence training (SCST) showed that captioning models can be optimized with policy-gradient learning using task-level rewards computed on full captions [5]. In SCST, the model's own inference-time output is used as a baseline, and sampled captions are encouraged when they obtain higher rewards than the baseline. This approach made it possible to optimize non-differentiable captioning metrics more directly than maximum-likelihood training. However, early applications of sequence-level optimization often focused on a single reward or on text-overlap-oriented metrics, which may be insufficient for medical image captioning where visual grounding and clinical concept consistency are also important.

A related direction has become central in modern language-model post-training. Reinforcement learning from human feedback and preference-based optimization methods refine supervised models using rewards or preferences defined over complete model responses [26]. More recent group-relative methods compare multiple sampled responses for the same input and compute relative advantages without requiring a separately trained value model. For example, GRPO uses group-relative normalization among generated outputs and has been used for post-training large language models in reasoning-oriented settings [6]. GDPO further extends this idea to multi-reward optimization by decoupling reward-specific contributions and normalizing heterogeneous reward signals before aggregation [7]. These properties are useful when the desired output quality is inherently multi-dimensional.

This multi-dimensionality is central to medical image captioning. A generated caption should recover the semantic content of the reference caption, preserve clinically meaningful terminology, remain compatible with the visual evidence, and avoid unsupported or incomplete clinical concepts. Optimizing a single captioning metric is therefore unlikely to capture all desired properties. In our previous ImageCLEFmedical Caption 2025 system, metric-aware selection was performed at inference time by generating multiple candidate captions and reranking them using biomedical text and image–text similarity signals [19]. Such reranking strategies can improve the final choice among available candidates, but they do not update the captioning model itself and are limited by the diversity and quality of the generated candidate pool.

Our work adopts a training-time sequence-level optimization perspective. After supervised fine-

tuning, we apply GDPO to refine the captioning policy using multiple reward signals computed on complete generated captions. Unlike inference-time reranking, this post-SFT optimization directly changes the model distribution toward captions that better satisfy the reward criteria. Unlike single-metric caption optimization, our formulation combines caption-to-caption, caption-to-image, and caption-to-CUI alignment signals. This makes reward balancing a central design issue, which we address through reward-decoupled advantage normalization and reward scheduling.

2.4. Evaluation Metrics and Clinical Concept Consistency

Medical image captioning evaluation is inherently multi-dimensional. A caption can be lexically similar to a reference while still being visually weakly grounded, and it can be visually plausible while omitting clinically important concepts. The ImageCLEFmedical Caption 2026 benchmark reflects this difficulty by evaluating generated captions with multiple complementary metrics grouped into relevance and factuality dimensions [3, 27]. This evaluation design motivates training and selection strategies that account for reference-caption similarity, image-caption compatibility, and clinical concept preservation rather than optimizing a single text-overlap metric.

Text-based metrics capture different aspects of reference-caption alignment. ROUGE measures lexical overlap and has been widely used for summarization and generation evaluation [9]. BERTScore compares contextual token representations from pretrained language models and is less dependent on exact surface-form matching than purely lexical metrics [8]. BLEURT is a learned evaluation metric trained to better correlate with human judgments of generated text quality [28]. These metrics are useful for measuring semantic and lexical similarity to reference captions, but they are not sufficient for medical image captioning because a generated caption may be textually plausible while remaining unsupported by the image or clinically incomplete.

Visual grounding is therefore another important evaluation axis. Image–text similarity metrics compare visual and textual representations to estimate whether a generated caption is compatible with the input image. Biomedical vision-language models such as BiomedCLIP provide domain-specific image–text representations learned from biomedical image–text pairs and can therefore be used as proxy models for medical image-caption compatibility [10]. In our training pipeline, this motivates the use of a BiomedCLIP-based image–text similarity reward. However, such similarity scores should be interpreted as coarse grounding signals rather than complete measures of clinical correctness, because embedding similarity may not capture fine-grained errors such as laterality mismatches, incorrect assertion status, or unsupported diagnostic interpretations.

Clinical factuality also requires concept-level and claim-level evaluation. The Unified Medical Language System (UMLS) provides a standardized biomedical concept vocabulary that supports concept-level matching across surface forms [12]. Tools such as MedCAT perform medical concept annotation and linking, enabling generated clinical text to be evaluated in terms of extracted biomedical concepts rather than only word overlap [11]. Biomedical language models such as BioBERT, PubMedBERT, and BiomedBERT further show that domain-specific pretraining improves biomedical and clinical text understanding, making them useful foundations for concept extraction and clinical text representation [29, 13, 14]. In this work, these resources motivate the use of a frozen biomedical caption-to-CUI extractor as a training-time proxy for clinical concept coverage.

Beyond concept overlap, factual consistency also involves whether the claims in a generated caption are supported by the reference or source information. AlignScore evaluates factual consistency through a unified alignment function and provides a complementary view of support-based factuality [30]. This distinction is important for medical captioning: a caption may mention many correct biomedical concepts while still making an unsupported claim about laterality, negation, relation, or diagnosis. Therefore, concept-level rewards and alignment-based factuality metrics should be viewed as related but not identical measures of clinical faithfulness.

The official evaluation metrics motivate our reward design, but our training-time rewards are not intended to exactly replicate the official evaluator. Instead, we use BERTScore Recall and ROUGE-1 as caption-to-caption alignment rewards, BiomedCLIP-based image–text similarity as a proxy for visual

grounding, and a frozen caption-to-CUI extractor as a proxy for clinical concept coverage. These reward components provide complementary signals for post-SFT GDPO optimization while acknowledging that strict claim-level factuality remains a challenging and only partially captured objective.

3. Method

Our system follows a two-stage training pipeline consisting of supervised fine-tuning (SFT) followed by post-SFT GDPO. The first stage adapts a vision-language backbone to the medical image captioning format using teacher-forcing likelihood training. The second stage further optimizes the SFT checkpoint with sequence-level reward signals computed on complete generated captions. The same overall framework is used for both the standard and synthetical caption prediction settings, while the reward scheduling strategy and final submission-level post-processing are selected separately for each setting.

The proposed SFT–GDPO pipeline is backbone-agnostic and can be instantiated with different pretrained vision-language checkpoints. In this work, we used three backbone checkpoints: `Qwen/Qwen3-VL-4B-Instruct`, a general-purpose instruction-tuned vision-language model from the Qwen-VL family [31, 21, 22, 32]; `OpenMed/Qwen3.5-2B-MedVL`, a medical vision-language checkpoint based on Qwen3.5-2B and fine-tuned on medical VQA data [25]; and `google/medgemma-4b-it`, a 4B multimodal instruction-tuned MedGemma checkpoint for medical text and image understanding [24]. For readability, we refer to these checkpoints as Qwen3-VL-4B, Qwen-MedVL-2B, and MedGemma-4B, respectively.

Figure 1 summarizes the proposed pipeline. Given an input medical image and a captioning prompt, the model is first trained to reproduce the reference caption through SFT. The resulting checkpoint is then used as the initial policy for GDPO, where multiple candidate captions are sampled for the same image and evaluated with a multi-alignment reward. The reward is organized around three alignment axes: caption-to-caption alignment, caption-to-image alignment, and caption-to-CUI alignment. These axes are implemented using four reward components: BERTScore Recall, ROUGE-1, a BiomedCLIP-based image–text similarity score, and a frozen biomedical caption-to-CUI extractor. During GDPO, reward-specific group-relative advantages are computed and then combined using schedule-dependent weights. Finally, selected checkpoints are used for deterministic decoding and submission-level candidate selection or fallback post-processing.

The following subsections describe the SFT stage, the GDPO formulation, the multi-objective reward components, the caption-to-CUI extractor, reward aggregation and scheduling, and the inference-time post-processing procedure.

3.1. Supervised Fine-Tuning

The first stage of our pipeline performs supervised fine-tuning on medical image–caption pairs. Given an input medical image I, a textual instruction context x, and a reference caption y^*, the captioning model is trained to generate y^* conditioned on (I, x). The model policy is denoted by π_θ, where θ represents the trainable parameters of the vision-language model.

We use a chat-style vision-language prompt format consisting of a system instruction, the input image, a user instruction, and an assistant response. The system and user instructions define the captioning task, while the assistant response contains the gold reference caption during training. The system prompt is:

```
You are a medical image captioning expert. Given a medical image,
provide a detailed and accurate caption describing the clinical
findings, imaging modality, and relevant anatomical structures.
```

The user prompt is:

```
<image>
Describe this medical image in detail.
```

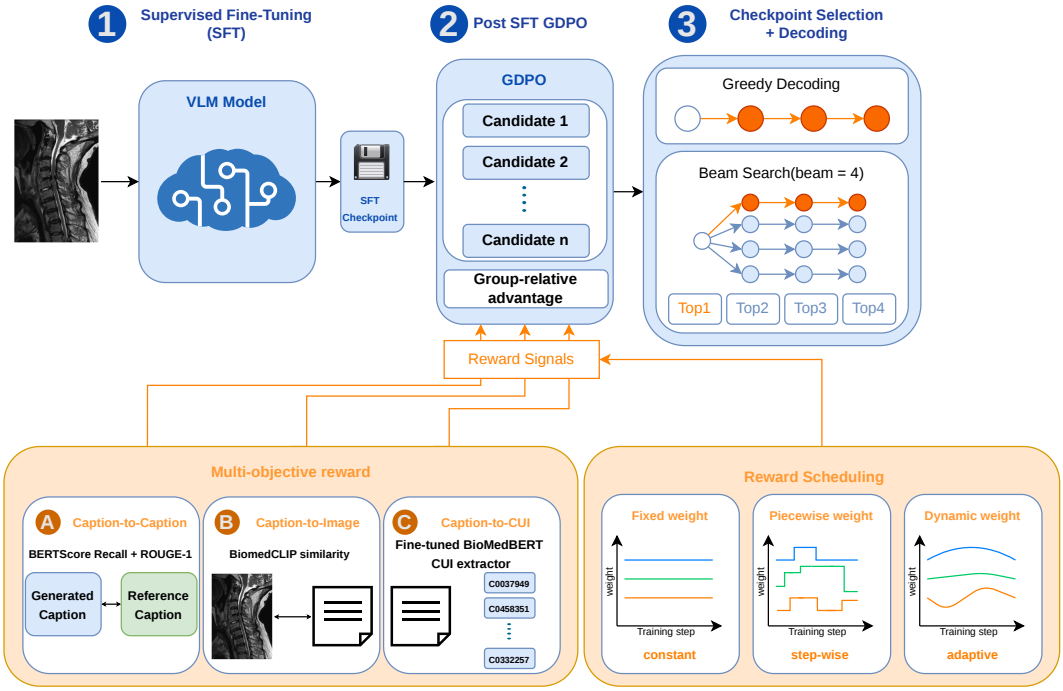


Figure 1: Overview of the proposed GDPO-based medical image captioning pipeline. After supervised fine-tuning, the model is optimized with post-SFT GDPO using multi-objective reward signals. The reward function is organized around three alignment axes: caption-to-caption alignment, caption-to-image alignment, and caption-to-CUI alignment. Reward scheduling controls the relative contribution of the reward components during GDPO training, and the final system uses checkpoint selection followed by greedy or beam-search decoding.; CC BY[Hongo et al. (2018)]

The same prompt format is used during SFT, GDPO rollout generation, and final inference. This design reduces prompt-format mismatch across training and evaluation stages and ensures that changes in model behavior are primarily due to optimization and decoding rather than different instruction formats.

We train the model with teacher-forcing next-token prediction. Let the reference caption be

$$y^* = (y_1^*, \ldots, y_T^*).$$

The SFT objective is defined as

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\sum_{t=1}^{T} \log \pi_\theta \left(y_t^* \mid y_{<t}^*, x, I\right), \tag{1}$$

where (x, I) denotes the textual instruction context and visual input. The loss is applied only to assistant caption tokens; tokens belonging to the system prompt, user prompt, and image placeholder are used as conditioning context and are masked from the training loss.

The purpose of SFT is to obtain a stable medical captioning policy before reinforcement learning. Through this stage, the model learns the caption style of the benchmark, modality-specific expressions, common anatomical terminology, and basic image–caption alignment. However, because the SFT objective optimizes token-level likelihood, it does not directly optimize complete-caption quality or the multi-dimensional evaluation criteria used in the benchmark. The SFT checkpoint therefore serves both as the supervised baseline and as the initial policy for the subsequent GDPO stage, where generated captions are further optimized using sequence-level reward signals.

3.2. Post-SFT GDPO

After supervised fine-tuning, the SFT checkpoint is used as the initial policy for post-SFT GDPO. This stage refines the captioning model using rewards computed on complete generated captions rather than relying only on token-level likelihood. This is important for medical image captioning because clinically relevant errors, such as omitted findings, weak visual grounding, unsupported clinical concepts, or incorrect anatomical descriptions, are usually properties of the full generated caption.

For each image–prompt input (I, x), the current policy generates a group of G candidate captions:

$$\{y_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot \mid I, x), \tag{2}$$

where $\pi_{\theta_{\text{old}}}$ denotes the rollout policy before the current update. Each candidate caption is evaluated by multiple reward components. For reward component k, we write

$$R_{i,k} = R_k(y_i; I, y^*, C^*), \tag{3}$$

where y^* is the reference caption and C^* is the gold CUI set. The reference caption and gold CUI set are used only for training-time reward computation and are not available during inference.

GDPO converts candidate-level rewards into group-relative advantages. In our setting, all candidates within a group are generated for the same image and prompt. Therefore, group-wise normalization compares alternative captions under the same visual input and reduces the effect of image-level reward-scale variation. For each reward component k, we compute

$$A_{i,k} = \frac{R_{i,k} - \mu_k}{\sigma_k + \epsilon}, \qquad \mu_k = \frac{1}{G}\sum_{j=1}^{G} R_{j,k}, \qquad \sigma_k = \sqrt{\frac{1}{G}\sum_{j=1}^{G}(R_{j,k} - \mu_k)^2}, \tag{4}$$

where ϵ is a small constant for numerical stability. If all candidates receive nearly identical scores for a reward component, the corresponding normalized advantages become small, and that component contributes little to the update for that group.

The reward-specific advantages are then combined using schedule-dependent weights. Let $w_k(t)$ denote the weight assigned to reward component k at training step t. The total caption-level advantage for candidate i is

$$A_i^{\text{total}}(t) = \sum_{k=1}^{K} w_k(t) A_{i,k}. \tag{5}$$

In our implementation, the reward components correspond to BERTScore Recall, ROUGE-1, caption-to-CUI Soft F1, and BiomedCLIP-based image–text similarity. The reward definitions are described in Section 3.3, and the scheduling strategies for $w_k(t)$ are described in Section 3.5.

The total advantage is used to update the captioning policy. Following GDPO [7], we use a clipped policy-gradient objective with a KL regularization term to keep the updated policy close to the SFT reference policy. For a generated caption $y_i = (y_{i,1}, \ldots, y_{i,T_i})$, define the token-level probability ratio

$$r_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t} \mid y_{i,<t}, x, I)}{\pi_{\theta_{\text{old}}}(y_{i,t} \mid y_{i,<t}, x, I)}. \tag{6}$$

The GDPO loss is written as

$$\begin{aligned}\mathcal{L}_{\text{GDPO}}(\theta) = -\frac{1}{G}\sum_{i=1}^{G}\frac{1}{T_i}\sum_{t=1}^{T_i}\min\left[r_{i,t}(\theta)A_i^{\text{total}},\ \text{clip}\left(r_{i,t}(\theta), 1-\epsilon_{\text{clip}}, 1+\epsilon_{\text{clip}}\right)A_i^{\text{total}}\right] \\ + \beta D_{\text{KL}}\left(\pi_\theta(\cdot \mid I, x) \,\|\, \pi_{\text{ref}}(\cdot \mid I, x)\right).\end{aligned} \tag{7}$$

where π_{ref} is the frozen SFT reference policy, ϵ_{clip} is the clipping parameter, and β controls the strength of the KL penalty. The caption-level advantage is shared across the generated response tokens of the same candidate caption. Intuitively, the objective increases the probability of captions with positive

group-relative advantages and suppresses captions with lower relative reward, while the clipping and KL terms prevent overly large policy updates.

This formulation is reward-decoupled because each reward dimension is normalized within the candidate group before aggregation. As a result, heterogeneous rewards with different scales, sparsity patterns, and noise levels can be combined without allowing a single raw reward magnitude to dominate the policy update. GDPO is therefore used as a post-SFT refinement stage: SFT provides a stable medical captioning policy, and GDPO further shifts the model distribution toward captions that better satisfy the multi-alignment reward criteria.

3.3. Multi-Objective Reward Components

The GDPO stage uses a multi-alignment reward designed to capture complementary aspects of medical caption quality. The reward is organized around three alignment axes: caption-to-caption alignment, caption-to-image alignment, and caption-to-CUI alignment. These axes are implemented with four reward components: BERTScore Recall, ROUGE-1 F-measure, caption-to-CUI Soft F1, and BiomedCLIP-based image–text similarity. For a generated caption y, input image I, reference caption y^*, and gold CUI set C^*, we define the reward vector as

$$\mathbf{R}(y; I, y^*, C^*) = [R_{\mathrm{BERT}}(y, y^*), R_{\mathrm{ROUGE}}(y, y^*), R_{\mathrm{CUI}}(y, C^*), R_{\mathrm{SIM}}(y, I)]\,. \tag{8}$$

Each reward component is first converted into a reward-specific group-relative advantage and the resulting advantages are combined using schedule-dependent weights, as described in Section 3.5. This design allows text-based reference alignment, visual grounding, and clinical concept coverage to contribute complementary training-time signals.

Caption-to-caption rewards. The caption-to-caption rewards measure how well the generated caption matches the reference caption at semantic and lexical levels. We use BERTScore Recall and ROUGE-1 F-measure:

$$R_{\mathrm{BERT}}(y, y^*) = \mathrm{BERTScore}_{\mathrm{Recall}}(y, y^*), \tag{9}$$

$$R_{\mathrm{ROUGE}}(y, y^*) = \mathrm{ROUGE\text{-}1}_{\mathrm{F1}}(y, y^*). \tag{10}$$

BERTScore Recall encourages recovery of semantic content from the reference caption, which is important for reducing clinically relevant omissions. ROUGE-1 F-measure complements this by encouraging lexical consistency with reference captions and common medical terminology. These rewards are useful for reference-caption alignment, but they do not by themselves guarantee that the generated caption is visually grounded or clinically complete.

Caption-to-image similarity reward. To encourage visual grounding, we use a BiomedCLIP-based image–text similarity reward. Let $\mathbf{v}_I$ denote the BiomedCLIP image embedding of image I, and let $\mathbf{v}_y$ denote the BiomedCLIP text embedding of generated caption y. The similarity reward is defined as

$$R_{\mathrm{SIM}}(y, I) = 2.5 \cdot \max\left(0, \frac{\mathbf{v}_I^\top \mathbf{v}_y}{\|\mathbf{v}_I\|_2 \|\mathbf{v}_y\|_2}\right). \tag{11}$$

This transformation clips negative cosine similarities to zero and scales positive similarities by 2.5. The reward is intended to discourage generic but plausible medical captions by favoring captions that are compatible with the visual input. However, this reward should be interpreted as a proxy for visual grounding rather than a complete measure of clinical correctness. In particular, embedding-based similarity may not capture fine-grained clinical errors such as laterality mismatches, incorrect assertion status, or unsupported diagnostic interpretations.

Caption-to-CUI reward. The caption-to-CUI reward encourages clinical concept consistency between the generated caption and the gold concept annotations. Because generated captions are free-form text while CUI annotations are discrete concept labels, we use a separately trained biomedical caption-to-CUI extractor f_ϕ to obtain soft concept predictions. The extractor is trained before GDPO and frozen during reward computation, as described in Section 3.4.

Let $\mathcal{C}_{\text{freq}}$ denote the retained CUI label space after rare-label filtering:

$$\mathcal{C}_{\text{freq}} = \{c \in \mathcal{C} \mid n_c \geq 20\}, \tag{12}$$

where n_c is the number of training captions assigned to concept c. Given a generated caption y, the extractor outputs a soft concept-probability vector

$$\hat{\mathbf{c}}(y) = f_\phi(y) \in [0,1]^{|\mathcal{C}_{\text{freq}}|}. \tag{13}$$

The gold CUI set C^* is converted into a binary vector $\mathbf{c}^*$ over the same retained label space. We compute a threshold-free Soft F1 reward as

$$R_{\text{CUI}}(y, C^*) = \text{SoftF1}\left(\hat{\mathbf{c}}(y), \mathbf{c}^*\right) = \frac{2\sum_{c\in\mathcal{C}_{\text{freq}}} \hat{c}_c(y) c_c^*}{\sum_{c\in\mathcal{C}_{\text{freq}}} \hat{c}_c(y) + \sum_{c\in\mathcal{C}_{\text{freq}}} c_c^* + \epsilon}, \tag{14}$$

where ϵ is a small constant for numerical stability. This threshold-free formulation provides a smooth concept-level reward signal and avoids sensitivity to a manually chosen decision threshold during GDPO.

The CUI reward is designed to encourage frequent clinical concept coverage, but it is still an imperfect proxy for clinical factuality. Rare-label filtering improves reward stability but may under-reward clinically important rare concepts. In addition, concept overlap does not fully represent assertion status, negation, laterality, relations, or complete clinical entailment. Therefore, R_{CUI} is treated as a bounded auxiliary reward for concept-level consistency rather than a complete factuality reward or an exact replica of the official UMLS Concept F1 metric.

3.4. Caption-to-CUI Extractor

Before the GDPO stage, we train a caption-to-CUI extractor to provide the caption-to-CUI reward described in Section 3.3. The purpose of this extractor is to map a free-form generated caption to a set of soft biomedical concept predictions. These predictions are then compared with the gold CUI annotations to compute a smooth concept-coverage reward during GDPO.

The extractor is initialized from `microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext` and fine-tuned as a multi-label classifier using reference captions and their associated gold CUI annotations [13, 14]. Given a caption, the model outputs a probability for each retained CUI label. Because the original CUI label space is long-tailed, we remove rare labels before training the extractor. Specifically, a CUI label is retained only if it appears in at least 20 training samples:

$$\mathcal{C}_{\text{freq}} = \{c \in \mathcal{C} \mid n_c \geq 20\}.$$

The same retained label space $\mathcal{C}_{\text{freq}}$ is used for extractor training, validation, and GDPO reward computation. This filtering reduces noisy and extremely sparse reward dimensions, although it also means that rare but clinically meaningful concepts may be under-represented in the CUI reward.

The extractor is trained independently before GDPO and then frozen during policy optimization. Freezing the extractor is important because it keeps the reward function fixed while the captioning policy changes. If the extractor were updated during GDPO, the reward model itself could adapt to the generated captions, making the reward signal unstable and harder to interpret. By using a fixed extractor, the captioning policy receives a stable concept-level reward that is separated from the generator being optimized.

Table 1
Validation performance of the BiomedBERT-based caption-to-CUI extractor after rare-label filtering. Threshold-dependent scores use a decision threshold of 0.90.

Dataset	Total CUIs	Retained CUIs	Removed CUIs	Micro-F1	Macro-F1	Jaccard	mAP	Thr.
Standard	2,646	1,345	1,301	0.9297	0.9013	0.8871	0.9390	0.90
Synthetical	2,174	1,389	785	0.9848	0.9494	0.9755	0.9677	0.90

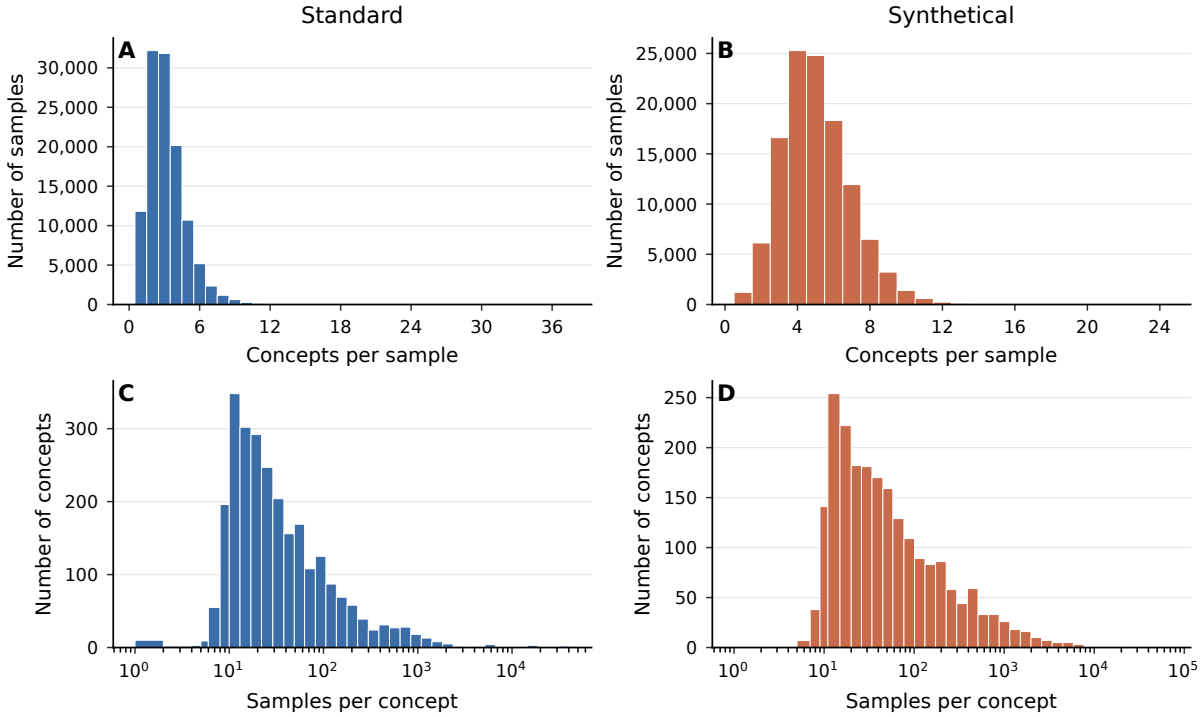


Figure 2: Distribution of CUI annotations before rare-label filtering. The upper panels show the number of assigned CUI concepts per training sample, and the lower panels show the number of training samples assigned to each CUI concept. Both standard and synthetical splits exhibit long-tailed CUI distributions, motivating the removal of labels that appear fewer than 20 times.

Table 1 reports the validation performance of the caption-to-CUI extractor after rare-label filtering. The extractor achieves high validation performance in both task settings, indicating that reference captions contain enough lexical and contextual information to recover frequent CUI labels. The synthetical setting shows higher validation scores than the standard setting, which is consistent with its denser and more regular concept annotations. Figure 2 further shows that both the standard and synthetical training splits contain long-tailed CUI distributions, supporting the use of rare-label filtering before extractor training.

Although the extractor performs well on validation captions, it should be interpreted as a reward-side proxy model rather than as a complete clinical factuality evaluator. First, extractor errors can propagate into the GDPO reward. Second, the filtered label space focuses on frequent concepts and does not fully cover rare clinical entities. Third, concept-level prediction does not represent all aspects of factuality, such as negation, assertion status, laterality, relations between findings and anatomy, or complete clinical entailment. For these reasons, the caption-to-CUI reward is used as a bounded auxiliary signal for frequent clinical concept coverage, complementing the caption-to-caption and caption-to-image reward components rather than replacing claim-level factuality evaluation.

3.5. Reward Aggregation and Scheduling

The reward components defined in Section 3.3 differ in scale, sparsity, noise level, and reliability. For example, text-based caption-to-caption rewards are dense and directly related to reference-caption similarity, whereas the caption-to-CUI reward depends on the accuracy of a separately trained concept extractor. The BiomedCLIP-based image–text similarity reward provides a visual-grounding signal, but it may be less sensitive to fine-grained clinical errors. Therefore, simply summing raw reward values can make optimization unstable or allow one reward dimension to dominate the update.

To address this issue, we first convert each reward component into a group-relative advantage, as defined in Equation (4). The final advantage for candidate i is then computed by a weighted sum of reward-specific advantages:

$$A_i^{\text{total}}(t) = w_{\text{BERT}}(t)A_{i,\text{BERT}} + w_{\text{ROUGE}}(t)A_{i,\text{ROUGE}} + w_{\text{CUI}}(t)A_{i,\text{CUI}} + w_{\text{SIM}}(t)A_{i,\text{SIM}}, \quad (15)$$

where $w_k(t)$ denotes the schedule-dependent weight for reward component k at training step t. Because aggregation is performed after reward-specific normalization, the policy update depends on the relative quality of candidate captions within each reward dimension rather than on raw reward magnitudes.

Fixed weighting. As a simple baseline, we use constant reward weights throughout GDPO training:

$$\mathbf{w}_{\text{fixed}} = [0.50,\ 0.50,\ 0.30,\ 0.30], \quad (16)$$

where the entries correspond to BERTScore Recall, ROUGE-1 F-measure, caption-to-CUI Soft F1, and BiomedCLIP-based image–text similarity, respectively. This configuration gives the largest total weight to caption-to-caption alignment while retaining auxiliary concept-coverage and visual-grounding signals. The total weight mass is 1.60. Although fixed weighting is simple and reproducible, it cannot adapt to training-stage differences or reward conflicts.

Piecewise scheduling. For the standard caption prediction setting, we use a manually designed piecewise schedule. The schedule was chosen after inspecting the behavior of the fixed-weight GDPO run. It keeps the initial fixed weights during the early stage of training, then gradually increases the weights of text alignment and CUI coverage while reducing the image–text similarity weight in the later stage. The initial, middle, and final weight vectors are

$$\mathbf{w}_0 = [0.50,\ 0.50,\ 0.30,\ 0.30],\ \ \mathbf{w}_{\text{mid}} = [0.55,\ 0.58,\ 0.30,\ 0.22],\ \ \mathbf{w}_{\text{final}} = [0.65,\ 0.70,\ 0.35,\ 0.15]. \quad (17)$$

The schedule is defined as

$$\mathbf{w}(t) = \begin{cases} \mathbf{w}_0, & t \le 1000, \\ \mathbf{w}_0 + (\mathbf{w}_{\text{mid}} - \mathbf{w}_0)\,\dfrac{t-1000}{800}, & 1000 < t \le 1800, \\ \mathbf{w}_{\text{mid}} + (\mathbf{w}_{\text{final}} - \mathbf{w}_{\text{mid}})\,\dfrac{t-1800}{800}, & 1800 < t \le 2600, \\ \mathbf{w}_{\text{final}}, & t > 2600. \end{cases} \quad (18)$$

This schedule keeps the image-similarity reward as an auxiliary grounding signal while placing greater emphasis on reference-caption alignment and CUI coverage as training progresses. The motivation is that visual compatibility is useful for discouraging generic captions in the early stage, whereas later updates should focus more strongly on recovering reference semantics and clinically relevant concepts.

Dynamic scheduling. For the synthetical caption prediction setting, we use an online dynamic scheduler based on reward-level advantage statistics. Dynamic scheduling is disabled during the first 800 steps, where the fixed weights in Equation (16) are used. From step 801 onward, the scheduler updates the next-step reward weights at every training step.

For each rollout group g, candidate i, and reward component k, we first compute the group-normalized advantage

$$A_{g,i,k} = \frac{R_{g,i,k} - \mu_{g,k}}{\sigma_{g,k} + \epsilon}, \qquad \epsilon = 10^{-8}. \tag{19}$$

For each reward component k, a consensus advantage is computed from the other reward dimensions by excluding k and taking the current-weighted average of the remaining normalized advantages. The scheduler then estimates three batch-level quantities. The first is *alignment*, defined as the Pearson correlation between the advantage of reward component k and the consensus advantage. The second is *signal*, defined as the fraction of rollout groups whose raw reward standard deviation exceeds 10^{-4}. The third is *risk*, defined as the fraction of top-quartile candidates under reward component k whose consensus score is below the median consensus score. These statistics measure whether a reward component agrees with the other rewards, provides non-degenerate variation, and avoids selecting candidates that conflict with the remaining reward dimensions.

The alignment, signal, and risk values are smoothed with an exponential moving average using smoothing coefficient $\alpha_m = 0.05$. The utility of reward component k is defined as

$$\mathrm{util}_k = \mathrm{align}_k \cdot \mathrm{signal}_k \cdot s_k - 0.5 \cdot \mathrm{risk}_k, \tag{20}$$

where

$$\mathbf{s} = [1.0,\ 1.0,\ 0.5,\ 0.3]$$

scales the utilities for BERTScore Recall, ROUGE-1 F-measure, caption-to-CUI Soft F1, and image–text similarity, respectively. The target weight is updated multiplicatively and then smoothed:

$$\widetilde{w}_k = \mathrm{clip}\left(w_k \exp(0.03\,\mathrm{util}_k), l_k, b_k\right), \qquad w_k^{\mathrm{new}} = 0.95 w_k + 0.05 \widetilde{w}_k. \tag{21}$$

The lower and upper bounds are

$$\mathbf{l} = [0.45,\ 0.45,\ 0.18,\ 0.08], \qquad \mathbf{b} = [0.75,\ 0.80,\ 0.36,\ 0.32].$$

After the multiplicative update and smoothing, the weight vector is projected to satisfy additional mass constraints: the total weight mass is constrained to 1.60, the text-reward mass $w_{\mathrm{BERT}} + w_{\mathrm{ROUGE}}$ is constrained to the interval $[1.00, 1.30]$, and the auxiliary-reward mass $w_{\mathrm{CUI}} + w_{\mathrm{SIM}}$ is constrained to the interval $[0.35, 0.60]$. After step 1701, the auxiliary upper bound is further reduced to 0.50. The newly computed weights are applied with a one-step delay: step t uses the current weights to compute advantages and stores the next weights, while step $t + 1$ applies the stored weights.

Interpretation. The two scheduling strategies are intended as data-regime-specific mechanisms for balancing heterogeneous reward signals, rather than as universally ranked alternatives. The standard and synthetical settings may differ in reward reliability, concept density, and reward conflict. The piecewise schedule encodes a manually chosen training-stage prior, whereas the dynamic scheduler adjusts weights online based on agreement among reward-specific advantages. As shown in Figure 3, both schedules shift weight toward caption-to-caption rewards and reduce the image–text similarity weight during later training, suggesting that visual grounding is most useful as an auxiliary signal while reference and concept alignment become more important for final caption quality.

3.6. Inference and Final Submission Post-Processing

After SFT and GDPO training, final predictions were generated from selected checkpoints using deterministic decoding. Greedy decoding was first used for checkpoint screening on the validation split. For selected checkpoints, we additionally applied beam-search decoding and retained multiple beam hypotheses. The top-ranked beam output was treated as the main single-decoder prediction, while the remaining beam hypotheses and checkpoint-wise greedy predictions were used to construct candidate pools for optional submission-level selection. Detailed decoding hyperparameters are provided in Appendix A.1.

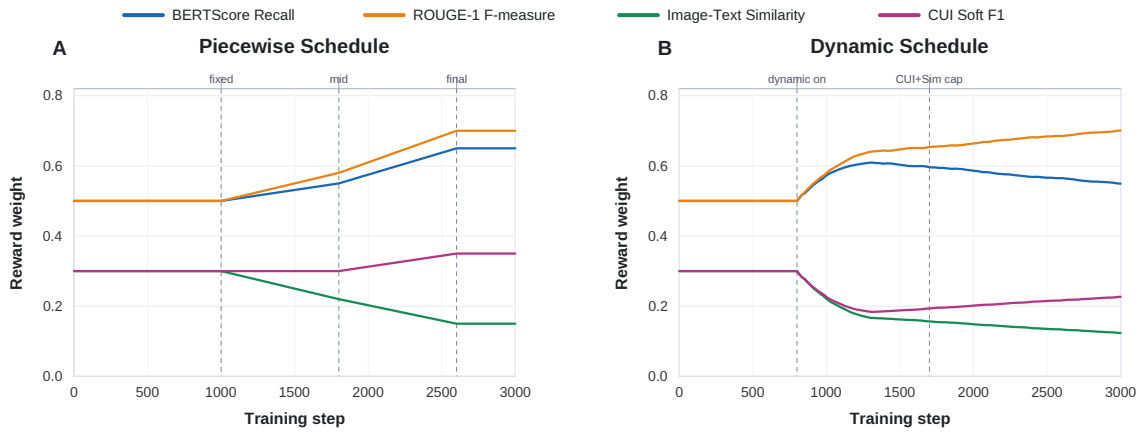


Figure 3: Reward-weight trajectories for the piecewise and dynamic scheduling strategies. The piecewise schedule follows predefined interpolation intervals, while the dynamic schedule updates weights online using advantage-alignment statistics under predefined constraints. Despite different update mechanisms, both schedules shift weight toward caption-to-caption rewards and reduce the image-text similarity weight during later training.

Candidate pool construction. For each image, the candidate pool consisted of predictions generated from selected checkpoints, decoding strategies, and model variants. In the standard track, the pool included greedy and beam-search outputs from fixed-weight and piecewise-scheduled Qwen-MedVL-2B GDPO checkpoints, together with selected MedGemma-4B candidates. In the synthetical track, the pool was constructed primarily from fixed-weight and dynamic Qwen-MedVL-2B GDPO checkpoints. These candidate pools were used only at inference time and did not introduce any additional training data or ground-truth information.

RadGraph-XL-guided candidate reranking. For the standard track, we evaluated an additional RadGraph-XL-guided re-boost variant [33]. This variant was used only for one of the submitted runs and was not used for the main synthetical track selection. The reranker selected one caption from the existing candidate set for each image and did not generate new text.

The reranking module first computed lightweight surface and clinical-coverage features, including modality terms, anatomical terms, disease-related terms, generic phrase indicators, and surface-quality penalties. It then used RadGraph-XL to extract clinical graph structures from candidate captions. For each image, the extracted structures from all candidates were aggregated into a candidate-consensus representation over observations, assertion status, anatomical entities, relations, and laterality. This consensus should be interpreted as a heuristic agreement signal among generated candidates, not as an external ground-truth label.

A candidate was rewarded when its extracted graph agreed with the candidate-level consensus and penalized when it introduced conflicting observations, assertion-status conflicts, laterality conflicts, generic wording, or low-quality surface forms. The final reranking score was computed as

$$\begin{aligned} S(y) = w_{\text{rank}} r(y) + w_{\text{concept}} S_{\text{concept}}(y) + w_{\text{consensus}} S_{\text{consensus}}(y) \\ + w_{\text{graph}} S_{\text{RG}}(y) - w_{\text{penalty}} P_{\text{quality}}(y), \end{aligned} \tag{22}$$

where $r(y)$ is the candidate rank index, with lower ranks corresponding to higher-priority base candidates. S_{concept} measures basic clinical concept coverage, $S_{\text{consensus}}$ measures agreement with the candidate-level textual consensus, S_{RG} is the RadGraph-XL graph-consensus score, and P_{quality} is a rule-based quality penalty. We used $w_{\text{rank}} = -0.05$, $w_{\text{concept}} = 0.8$, $w_{\text{consensus}} = 0.7$, $w_{\text{graph}} = 0.5$, and $w_{\text{penalty}} = 0.8$.

Rule-based fallback post-processing. For selected final submissions, we used a conservative rule-based fallback strategy to avoid submitting obviously malformed captions. The fallback module was designed not to rewrite fluent predictions. For each image, the base caption was first evaluated by a deterministic quality gate. If the base caption passed the gate, it was kept unchanged. If it failed, fallback candidates were scanned in a predefined order, and the first candidate that passed the same quality gate was selected. If no fallback candidate passed the gate, the original base caption was retained. This left-to-right design makes the procedure deterministic and preserves high-quality base predictions whenever possible.

The quality gate flagged captions with empty or invalid literal values, broken encoding artifacts, insufficient length, too few words, low alphabetic-character ratio, low ASCII-letter ratio, excessive punctuation, consecutive repeated words, overrepresented repeated words, and repeated 2-grams or 3-grams. The main thresholds were: at least 6 words, at least 24 characters, alphabetic-character ratio ≥ 0.45, ASCII-letter ratio ≥ 0.90, punctuation ratio ≤ 0.28, single-word frequency ratio ≤ 0.35, and repeated n-gram count below 3.

The base and fallback order used for each track is summarized in Table 2. All entries in the fallback selector were Qwen-MedVL-2B variants. The fallback procedure should be interpreted as a submission-level robustness step rather than as a controlled ablation of the training method. It was intended to correct rare formatting or degeneration failures while leaving the base model prediction unchanged in normal cases.

Table 2
Base and fallback order used for rule-based post-processing. All entries in this fallback selector are Qwen-MedVL-2B variants; "B4 #k" denotes the rank-k caption from beam-search decoding with beam size 4, and "FBk" denotes the k-th fallback candidate.

Track	Base	FB1	FB2	FB3	FB4
Standard	fixed, B4 #1	piecewise, B4 #1	fixed, greedy@2750	piecewise, B4 #2	fixed, B4 #2
Synthetical	dynamic, greedy	fixed, B4 #1	dynamic, B4 #2	fixed, B4 #2	fixed, greedy

4. Experiments

4.1. Experimental Settings

Task setting. We evaluated our system on the ImageCLEFmedical Caption 2026 caption prediction subtasks for both the standard and synthetical task families [3, 17]. The challenge defines the two task families separately and instructs participants not to mix the corresponding datasets. We therefore trained, validated, selected checkpoints, and prepared submissions independently for the two settings.

The standard track is based on the real-image split, which contains 97,364 training captions, 19,240 validation captions, and 15,249 test images. The synthetical track is based on the synthetic-image split, which contains 97,222 training captions, 19,239 validation captions, and 15,262 test images. The standard and synthetical datasets were kept separate throughout supervised fine-tuning, GDPO training, checkpoint selection, decoding, and submission-level post-processing.

Backbone usage and model selection. All three backbone checkpoints defined in Section 3 were trained and evaluated under the same two-stage SFT–GDPO framework. Qwen3-VL-4B served as a strong general-purpose multimodal comparison backbone. Qwen-MedVL-2B was used as the primary submission backbone because it produced the strongest validation candidates among the evaluated checkpoints despite its smaller parameter count. MedGemma-4B was retained as a medical-domain comparison backbone and as an additional source of candidate captions for selected standard-track submissions.

Table 3
Main official submission variants and run IDs used in the result tables. Fallback denotes conservative rule-based fallback post-processing, and RG-XL rerank denotes RadGraph-XL-guided candidate reranking.

Track	ID	Backbone / source	Training and selection
Standard	STD-1	Qwen-MedVL-2B	Fixed-weight GDPO, no post-processing
	STD-2	Qwen-MedVL-2B	Piecewise-scheduled GDPO, no post-processing
	STD-3	MedGemma-4B	Fixed-weight GDPO, no post-processing
	STD-4	Qwen-MedVL-2B	Fixed-weight GDPO with fallback
	STD-5	Qwen-MedVL-2B + MedGemma-4B	Fixed-GDPO candidate pool with RG-XL rerank
Synthetical	SYN-1	Qwen-MedVL-2B	Fixed-weight GDPO, no post-processing
	SYN-2	Qwen-MedVL-2B	Dynamic-scheduled GDPO, no post-processing
	SYN-3	MedGemma-4B	Fixed-weight GDPO, no post-processing
	SYN-4	Qwen-MedVL-2B	Dynamic-scheduled GDPO with fallback
	SYN-5	Qwen-MedVL-2B	Fixed-GDPO candidate pool with fallback

Training and inference settings. All systems followed the same broad pipeline: supervised fine-tuning, post-SFT GDPO, checkpoint selection, deterministic decoding, and optional submission-level post-processing. Unless otherwise stated, the same training and inference settings were used across the standard and synthetical tracks, except for track-specific reward scheduling and final submission choices. The standard scheduled run used the piecewise reward schedule described in Section 3.5, whereas the synthetical scheduled run used the dynamic reward schedule. Detailed hyperparameters for SFT, GDPO rollout generation, final inference, and the training environment are provided in Appendix A.1.

Submission variants. Table 3 summarizes the official submission variants and assigns compact run IDs used in the result tables. Runs without post-processing use the top-ranked beam-search output from the corresponding trained model unless otherwise specified. Fallback variants apply the conservative rule-based selector described in Section 3.6; the base prediction is retained unless it fails the deterministic quality gate. The RadGraph-XL re-boost variant uses a mixed candidate pool from Qwen-MedVL-2B and MedGemma-4B checkpoints and applies the candidate reranking procedure described in Section 3.6.

These variants should be interpreted as official submission configurations rather than exhaustive controlled ablations. In particular, fallback post-processing and RadGraph-XL-guided reranking were applied only to selected submitted runs. Controlled validation comparisons between SFT and GDPO are reported separately in Section 4.5.

4.2. Evaluation Metrics

The official ImageCLEFmedical Caption 2026 evaluation reports six individual metrics grouped into two higher-level dimensions: relevance and factuality [3, 17, 27]. Relevance measures whether a generated caption is similar to the reference caption and compatible with the input image. Factuality measures whether the generated caption preserves clinically meaningful concepts and contains information supported by the reference information.

The relevance score is computed as the average of four metrics: BERTScore Recall with inverse-document-frequency weighting, ROUGE-1 F-measure, BLEURT, and the official image–caption similarity score [8, 9, 28]. We denote these metrics by M_{BERT}, M_{ROUGE}, M_{BLEURT}, and M_{SIM}, respectively. The relevance score is therefore

$$\mathrm{Relevance} = \frac{1}{4}\left(M_{\mathrm{BERT}} + M_{\mathrm{ROUGE}} + M_{\mathrm{BLEURT}} + M_{\mathrm{SIM}}\right). \tag{23}$$

The factuality score is computed as the average of two metrics: UMLS Concept F1 and AlignScore [12, 30]. UMLS Concept F1 evaluates concept-level overlap between generated and reference captions using biomedical concept annotations, while AlignScore estimates whether the generated caption is

Table 4
Official results for standard caption prediction. Run IDs are defined in Table 3. Best values among our submitted runs are shown in bold.

ID	Overall	Rel.	Fact.	BERT	ROUGE	Sim.	BLEURT	UMLS F1	Align
STD-1	0.3753	0.5784	0.1722	0.6247	0.3231	**1.0104**	0.3554	0.2164	0.1281
STD-2	0.3747	**0.5794**	0.1700	**0.6266**	**0.3250**	1.0103	**0.3557**	0.2137	0.1264
STD-3	0.3730	0.5748	0.1712	0.6243	0.3194	1.0097	0.3459	0.2145	0.1280
STD-4	**0.3755**	0.5786	0.1724	0.6250	0.3234	**1.0104**	0.3555	0.2160	0.1287
STD-5	0.3754	0.5749	**0.1759**	0.6258	0.3216	1.0055	0.3467	**0.2193**	**0.1326**

supported by the reference information. Denoting these metrics by M_{UMLS} and M_{Align}, the factuality score is

$$\mathrm{Factuality} = \frac{1}{2}\left(M_{\mathrm{UMLS}} + M_{\mathrm{Align}}\right). \tag{24}$$

The official overall score is then computed as the average of relevance and factuality:

$$\mathrm{Overall} = \frac{1}{2}\left(\mathrm{Relevance} + \mathrm{Factuality}\right). \tag{25}$$

Thus, the overall score is not an unweighted average of the six individual metrics. Instead, the four relevance metrics jointly contribute one half of the overall score, and the two factuality metrics jointly contribute the other half.

In the result tables, we report the official overall, relevance, and factuality scores together with the six individual metrics. All metrics are reported in the official score direction, where higher values indicate better performance. The training-time rewards used in our GDPO stage are motivated by the same evaluation dimensions, but they should not be interpreted as exact replicas of the official evaluator. In particular, our BiomedCLIP-based similarity reward is a proxy for visual grounding and may differ from the official image–caption similarity metric.

4.3. Main Results on Standard Caption Prediction

Table 4 reports the official results for the standard Caption Prediction track using the run IDs defined in Table 3. Our best official standard-track submission was STD-4, the fixed-weight Qwen-MedVL-2B GDPO system with conservative rule-based fallback post-processing. It achieved an overall score of 0.3755 and ranked first by overall score on the official final leaderboard.

The Qwen-MedVL-2B variants formed a strong group of standard-track submissions. STD-1, the fixed-weight GDPO run without post-processing, already achieved an overall score of 0.3753. STD-4 improved the overall score slightly to 0.3755 after applying the fallback selector. Because the absolute difference between STD-1 and STD-4 is small, this comparison should be interpreted as a submission-level result rather than as evidence that fallback post-processing alone substantially improves caption quality. The main role of the fallback selector was to improve robustness by replacing obviously malformed captions while preserving the base prediction whenever it passed the quality gate.

STD-2, the piecewise-scheduled Qwen-MedVL-2B run, achieved the highest relevance score among our standard-track submissions. It also obtained the highest BERTScore Recall, ROUGE-1 F-measure, and BLEURT scores. However, its factuality score was lower than that of STD-4, which reduced its overall score despite stronger relevance-oriented performance. This result suggests that piecewise scheduling improved reference-caption alignment metrics, but did not uniformly improve the factuality-oriented metrics in the standard setting.

STD-5, the RadGraph-XL-guided re-boost variant, showed the opposite trade-off. It achieved the highest factuality score, UMLS Concept F1, and AlignScore among our standard-track submissions, indicating that graph-guided candidate reranking favored captions with stronger concept-level and support-based consistency. However, this improvement came with a lower relevance score, mainly due to decreases in image–caption similarity and BLEURT. As a result, STD-5 achieved an overall

Table 5
Official results for caption prediction synthetical. Run IDs are defined in Table 3. Best values among our submitted runs are shown in bold.

ID	Overall	Rel.	Fact.	BERT	ROUGE	Sim.	BLEURT	UMLS F1	Align
SYN-1	0.5838	**0.7189**	0.4488	0.7576	0.6342	**0.9771**	**0.5068**	0.5301	0.3674
SYN-2	0.5839	0.7182	**0.4496**	0.7577	0.6352	0.9737	0.5062	**0.5332**	0.3659
SYN-3	0.5736	0.7106	0.4367	0.7486	0.6227	0.9709	0.5001	0.5185	0.3549
SYN-4	**0.5840**	0.7183	**0.4496**	**0.7580**	**0.6353**	0.9737	0.5063	**0.5332**	0.3660
SYN-5	0.5839	**0.7189**	0.4488	0.7576	0.6343	0.9770	**0.5068**	0.5301	**0.3675**

score of 0.3754, slightly below STD-4. This illustrates the trade-off between relevance-oriented and factuality-oriented metrics under the official overall scoring scheme.

STD-3, the MedGemma-4B-based run, was competitive but remained below the Qwen-MedVL-2B variants in overall score. This is consistent with our validation analysis, where Qwen-MedVL-2B produced the strongest candidates among the evaluated backbones. Overall, the standard-track results suggest that the best official performance was obtained by combining a medical-domain-adapted Qwen-MedVL-2B backbone, fixed-weight GDPO, beam-search-based decoding, and a conservative fallback selector. At the same time, the small differences among the top submissions indicate that reward scheduling, reranking, and fallback post-processing should be interpreted as system-level design choices rather than isolated controlled ablations.

4.4. Main Results on Caption Prediction Synthetical

Table 5 reports the official results for the Caption Prediction Synthetical track using the run IDs defined in Table 3. Our best synthetical-track submission was SYN-4, the dynamic-scheduled Qwen-MedVL-2B GDPO system with conservative rule-based fallback post-processing. It achieved an overall score of 0.5840 and ranked first by overall score on the official final leaderboard.

The Qwen-MedVL-2B submissions were numerically very close in the synthetical track. SYN-4 achieved the highest overall score, but the differences among SYN-1, SYN-2, SYN-4, and SYN-5 were small. Therefore, these results should be interpreted as system-level submission comparisons rather than as isolated evidence that one scheduling or post-processing component dominates the others.

SYN-2, the dynamic-scheduled Qwen-MedVL-2B run without fallback post-processing, achieved an overall score of 0.5839, nearly matching the final SYN-4 submission. It also matched SYN-4 in factuality and UMLS Concept F1. This suggests that dynamic reward scheduling produced a strong base system for the synthetical setting, and that the fallback selector mainly acted as a conservative robustness mechanism rather than a major source of improvement.

SYN-4 achieved the best BERTScore Recall and ROUGE-1 F-measure among our synthetical submissions, while maintaining the highest factuality score. This combination led to the best overall score under the official evaluation scheme. However, SYN-1 and SYN-5 achieved slightly higher relevance scores, primarily because of stronger image–caption similarity and BLEURT scores. SYN-5 also obtained the highest AlignScore among our synthetical submissions. These metric patterns indicate that the synthetical track involved small trade-offs between reference-caption alignment, image–caption similarity, and support-based factuality.

SYN-3, the MedGemma-4B-based run, was competitive but remained below the Qwen-MedVL-2B variants by approximately 0.010 overall. This is consistent with our validation results, where Qwen-MedVL-2B produced the strongest synthetical captioning candidates among the evaluated backbones. Overall, the synthetical results suggest that medical-domain-adapted Qwen-MedVL-2B backbones, GDPO-based post-SFT optimization, and deterministic candidate selection provided a robust combination for the synthetic-image setting. At the same time, the very small differences among the top Qwen-based submissions indicate that the effects of dynamic scheduling and fallback post-processing should be interpreted cautiously and in conjunction with the validation analysis in Section 4.5.

Table 6
Internal validation comparison between SFT and fixed-weight GDPO on the standard and synthetical splits. Track abbreviations: Std. = Standard and Syn. = Synthetical. Backbone abbreviations: Qwen3 = Qwen3-VL-4B, Gemma = MedGemma-4B, and MedVL = Qwen-MedVL-2B. Metric abbreviations: Ovr. = Overall, R-1 = ROUGE-1, and BLT = BLEURT.

Track	Stage	Backbone	Ovr.	Rel.	Fact.	BERT	R-1	Sim.	BLT	UMLS	Align
Std.	SFT	Qwen3	0.3484	0.5373	0.1594	0.5929	0.2798	0.9488	0.3277	0.1819	0.1369
		Gemma	0.3499	0.5409	0.1589	0.5953	0.2787	0.9605	0.3293	0.1840	0.1339
		MedVL	0.3526	0.5443	0.1610	0.5981	0.2821	0.9665	0.3303	0.1857	0.1363
	GDPO	Qwen3	0.3607	0.5656	0.1558	0.6232	0.3094	0.9842	0.3454	0.2057	0.1059
		Gemma	0.3649	0.5715	0.1582	0.6238	0.3078	1.0102	0.3445	0.2017	0.1147
		MedVL	0.3660	0.5770	0.1550	0.6335	0.3100	1.0118	0.3527	0.2023	0.1076
Syn.	SFT	Qwen3	0.5431	0.6870	0.3992	0.7242	0.5919	0.9483	0.4838	0.4863	0.3121
		Gemma	0.5583	0.6973	0.4192	0.7294	0.6067	0.9596	0.4935	0.5019	0.3366
		MedVL	0.5633	0.7019	0.4246	0.7362	0.6148	0.9602	0.4964	0.5106	0.3387
	GDPO	Qwen3	0.5450	0.6878	0.4023	0.7258	0.5930	0.9467	0.4855	0.4859	0.3186
		Gemma	0.5690	0.7071	0.4309	0.7407	0.6189	0.9712	0.4978	0.5173	0.3444
		MedVL	0.5738	0.7129	0.4347	0.7479	0.6252	0.9772	0.5012	0.5209	0.3485

4.5. Validation-Based Analysis and Candidate Selection

In addition to the official test results, we used the released validation splits to analyze the effect of post-SFT optimization and to guide checkpoint, decoding, and candidate-pool selection. The validation results were used for within-split model selection and relative comparison. Their absolute values are not directly comparable to the official test scores because the validation and test splits may differ in caption distributions, image composition, and evaluation-time candidate selection procedures.

Table 6 compares supervised fine-tuning and fixed-weight GDPO on the standard and synthetical validation splits. Across all evaluated backbones and both task settings, GDPO improved the overall validation score compared with the corresponding SFT checkpoint. On the standard split, Qwen-MedVL-2B improved from 0.3526 after SFT to 0.3660 after GDPO. On the synthetical split, Qwen-MedVL-2B improved from 0.5633 to 0.5738. Similar improvements were also observed for Qwen3-VL-4B and MedGemma-4B. These results support the use of post-SFT sequence-level optimization for improving benchmark-oriented caption quality.

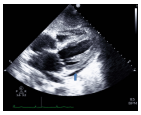
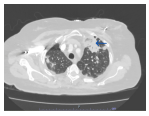
The metric-level changes were not uniform across task settings. On the standard split, GDPO substantially improved relevance-oriented metrics and UMLS Concept F1, but AlignScore decreased for all three backbones. For Qwen-MedVL-2B, AlignScore decreased from 0.1363 after SFT to 0.1076 after GDPO, even though the overall score increased from 0.3526 to 0.3660. This suggests that the reward family improved reference-caption similarity and concept coverage, but did not necessarily improve strict support-based factuality in the standard setting.

In contrast, on the synthetical split, GDPO improved both overall performance and AlignScore for all three backbones. For Qwen-MedVL-2B, AlignScore increased from 0.3387 after SFT to 0.3485 after GDPO. This difference between the standard and synthetical settings suggests that the interaction between reward optimization and alignment-style factuality depends on the data regime. In particular, denser or more regular concept supervision in the synthetical split may make the reward signals more consistent with factuality-oriented metrics.

The validation results also informed backbone and decoding choices. First, medical-domain or medical-adapted models outperformed the larger general-purpose Qwen3-VL-4B candidate. Qwen-MedVL-2B achieved the strongest validation results among the evaluated backbones after GDPO in both settings, despite being smaller than Qwen3-VL-4B. MedGemma-4B was also competitive, but its strongest validation scores remained below those of Qwen-MedVL-2B. These observations motivated the use of Qwen-MedVL-2B as the main submitted backbone, with MedGemma-4B retained as an additional candidate source for selected standard-track submissions.

Second, validation-based checkpoint screening supported the use of beam-search-based candidate

Table 7
Qualitative examples comparing reference and generated captions from validation predictions. The first example illustrates a well-grounded generation, whereas the second example shows a weakly grounded generation with fine-grained factual errors.

Image / case	Reference caption	Generated caption
 **Well-grounded case** CC BY[Lamour et al. (2024)]	The initial transthoracic echocardiogram on admission shows a small pericardial effusion (blue arrow) on the parasternal long-axis view.	Transthoracic echocardiogram in the parasternal long axis view showing a small pericardial effusion (blue arrow).
 **Weakly grounded case** CC BY[Muacevic et al. (2024)]	Computed tomography (CT) scan of the chest (image 2). The blue arrow is pointing to a large, hyperlucent lesion in the right upper lobe, likely septic emboli.	Computed tomography (CT) scan of the chest with contrast (axial view) of the lung showing a cavitary lesion in the left upper lobe (blue arrow).

generation. Greedy decoding was useful for efficient checkpoint screening, while beam-search decoding provided stronger candidate captions for final submission construction. In particular, Qwen-MedVL-2B beam-search hypotheses formed the strongest candidate pools in both standard and synthetical validation experiments. These results motivated our final emphasis on medical-domain initialization, post-SFT GDPO, selected Qwen-MedVL-2B checkpoints, and deterministic beam-search-based candidate generation.

Detailed checkpoint- and candidate-level validation results are provided in Appendix A.2. In brief, those additional results summarize the representative internal validation experiments used to select checkpoints, compare decoding variants, and construct candidate pools for the final submissions. They provide the empirical basis for the practical submission choices summarized above.

Overall, the validation analysis provides stronger evidence for the effectiveness of GDPO than the small differences among the final official submission variants alone. The official leaderboard results reflect the combined effects of model selection, reward scheduling, decoding, candidate selection, and fallback or reranking procedures, whereas the validation comparison in Table 6 directly contrasts SFT and fixed-weight GDPO under matched backbone and split conditions.

4.6. Qualitative Analysis

Table 7 presents representative qualitative examples from validation predictions. The examples are intended to illustrate typical strengths and remaining failure modes of the proposed system rather than to provide an exhaustive error analysis.

In the well-grounded example, the generated caption preserves the clinically important content of the reference caption. It correctly identifies the imaging modality as transthoracic echocardiography, describes the parasternal long-axis view, and retains the key finding of a small pericardial effusion. Although the generated caption is shorter and does not exactly reproduce the reference wording, it remains semantically faithful and visually grounded with respect to the main clinical observation.

The weakly grounded example illustrates a more challenging failure mode. The generated caption remains fluent and medically plausible, and it correctly recognizes the image as a chest CT showing a lung lesion. However, it introduces a clinically important laterality error by describing the lesion as being in the left upper lobe, whereas the reference caption specifies the right upper lobe. It also adds the phrase “with contrast,” which is not supported by the reference caption, and omits the reference-supported interpretation of likely septic emboli. These errors are clinically meaningful because laterality, contrast status, and diagnostic interpretation can change the factual content of a medical caption.

These examples show that the proposed reward design can improve reference-style caption generation

at the level of modality, view, anatomy, and major findings. However, they also indicate that fine-grained claim-level factuality remains difficult to control. In particular, BERTScore, ROUGE, image–text similarity, and CUI coverage can reward captions that are broadly relevant while still missing errors in laterality, assertion status, imaging protocol, or diagnostic relations. This observation is consistent with the quantitative results in Section 4.5, where GDPO improved overall validation performance but showed mixed behavior on AlignScore in the standard setting.

A more systematic error analysis suggests several directions for future improvement. Laterality errors may require relation-aware or graph-based rewards that explicitly model anatomical side and finding–anatomy relations. Unsupported imaging-protocol claims, such as adding contrast status when it is not stated or visually inferable, may require stricter entailment-style constraints. Omitted diagnostic interpretations may benefit from reward signals that distinguish concept coverage from complete clinical claim support. These remaining failure modes motivate future work on AlignScore-aware reward design, RadGraph-style relation rewards, and constraint-aware candidate selection.

5. Limitations and Future Work

This work has several limitations. First, the caption-to-CUI reward depends on a separately trained biomedical concept extractor that is frozen during GDPO. Although this design provides a stable training-time proxy for clinical concept coverage, extractor errors can propagate into the reward signal. In addition, the rare-label filtering step improves reward stability but removes infrequent CUI labels from the reward space. As a result, the CUI reward mainly reflects coverage of frequent biomedical concepts and may under-reward rare but clinically important findings.

Second, the BiomedCLIP-based image–text similarity reward is only a proxy for visual grounding. Embedding-based similarity can encourage coarse compatibility between an image and a caption, but it may not capture fine-grained clinical errors such as laterality mismatches, incorrect assertion status, unsupported contrast information, or erroneous finding–anatomy relations. Therefore, a caption can receive a reasonable image–text similarity reward while still containing clinically meaningful factual errors.

Third, reward scheduling introduces additional design choices. The piecewise schedule requires manually selected intervals and target weights, while the dynamic schedule depends on controller hyperparameters, smoothing factors, utility definitions, and weight-mass constraints. Although both schedules were useful in our system-level submissions, the present results do not establish a universal ranking between fixed, piecewise, and dynamic weighting strategies. The best schedule may depend on the data regime, reward reliability, concept density, and degree of conflict among reward components.

Another limitation is that the final leaderboard scores reflect the combined effect of multiple components, including backbone selection, supervised fine-tuning, GDPO, reward scheduling, checkpoint selection, beam-search decoding, candidate-pool construction, reranking, and fallback post-processing. Because reranking and fallback were applied only to selected official submission variants, the official submission comparisons should be interpreted as system-level results rather than exhaustive controlled ablations. In particular, the small score differences among several top submissions indicate that the isolated contribution of each inference-time component should be analyzed more systematically.

Finally, strict claim-level factuality remains a challenging and metric-sensitive direction. BERTScore, ROUGE, and BLEURT mainly measure reference-caption similarity, while the CUI reward encourages concept coverage without fully modeling negation, assertion status, laterality, relations, or complete clinical entailment [8, 9, 28]. In contrast, alignment-based factuality metrics such as AlignScore evaluate whether the information stated in a generated caption is supported by the reference or source information [30]. This difference helps explain why GDPO can improve overall performance through relevance-oriented and concept-coverage gains while showing mixed behavior on AlignScore, especially in the standard setting.

Future work will focus on four directions. First, we will conduct more systematic ablation studies under matched backbone, decoding, and post-processing conditions to isolate the effects of GDPO,

reward scheduling, candidate reranking, and fallback selection. Second, we will investigate relation-aware and graph-based rewards that explicitly model anatomical relations, laterality, assertion status, and finding–anatomy links. Third, we will explore AlignScore-aware or entailment-aware reward design to improve support-based factual consistency without degrading relevance-oriented metrics. Fourth, we will consider Pareto-aware reward scheduling and constraint-aware candidate selection to better balance relevance, visual grounding, concept coverage, and claim-level factuality. These extensions are expected to make post-SFT optimization more robust for clinically grounded medical image captioning beyond the current ImageCLEFmedical Caption setting.

6. Conclusion

We presented our ImageCLEFmedical Caption 2026 system for clinically grounded medical image captioning. The proposed system combines supervised fine-tuning with post-SFT GDPO and uses a multi-alignment reward organized around caption-to-caption, caption-to-image, and caption-to-CUI alignment. These alignment axes are implemented through BERTScore Recall, ROUGE-1, BiomedCLIP-based image–text similarity, and a frozen biomedical caption-to-CUI extractor. We also investigated reward scheduling strategies to balance heterogeneous reward components during GDPO training, using a piecewise schedule for the standard setting and a dynamic schedule for the synthetical setting.

Internal validation results showed that post-SFT GDPO improved overall validation performance over supervised fine-tuning across the evaluated backbones and task settings. The validation analysis also supported the importance of medical-domain adaptation: Qwen-MedVL-2B produced the strongest validation candidates among the evaluated backbones, despite being smaller than the general-purpose Qwen3-VL-4B model. These results suggest that medical-domain initialization and sequence-level post-SFT optimization are useful for improving benchmark-oriented medical caption generation.

Our best standard-track submission achieved an official overall score of 0.3755, and our best synthetical-track submission achieved an official overall score of 0.5840. Both systems ranked first by overall score on the official final leaderboards. These final results reflect the combined effects of backbone selection, supervised fine-tuning, GDPO, reward scheduling, checkpoint selection, beam-search-based decoding, and conservative submission-level post-processing. Therefore, they should be interpreted as system-level results rather than as isolated evidence for any single component.

Overall, the findings suggest that medical-domain vision-language backbones, multi-objective post-SFT optimization, and carefully controlled inference-time selection form an effective framework for competitive medical image captioning under the ImageCLEFmedical Caption 2026 evaluation protocol. At the same time, strict claim-level factuality remains an important open challenge. Future work will focus on relation-aware clinical rewards, AlignScore- or entailment-aware optimization, and more systematic ablations to better balance relevance, visual grounding, concept coverage, and support-based factual consistency.

Acknowledgments

We thank the ImageCLEFmedical Caption 2026 organizers for providing the benchmark, evaluation framework, and official result sheets. This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (grant number: RS-2024-00360176), and in part by a grant of the Korea Health Technology RD Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health Welfare, Republic of Korea (grant number: RS-2025-02222326).

Declaration on Generative AI

After completing the initial manuscript, the authors used ChatGPT solely for language editing, grammar correction, sentence polishing, and LaTeX formatting assistance. No generative AI tools were used to

develop the scientific content, conduct analyses, interpret results, or draw conclusions. All scientific content, analyses, interpretations, and conclusions were developed and verified by the authors. The authors reviewed and edited all AI-assisted outputs and take full responsibility for the final manuscript.

References

[1] D. R. Beddiar, M. Oussalah, T. Seppänen, R. Jennane, Acapmed: Automatic captioning for medical imaging, Applied Sciences 12 (2022) 11092.

[2] B. Ionescu, H. Müller, D. Stanciu, A. Radu, R. Bolborici, M. Negru, A. Ene, V. Vasilescu, A.-A. Nicolae, L. Ştefan, M. Constantin, M. Dogariu, A. Andrei, H. Damm, T. M. G. Pakull, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, R. Brüngel, L. Reinartz, H. Schäfer, C. S. Schmidt, B. Bracke, P. Nath, B. Eryılmaz, M. Hjuler, D. Fabre, C. Lemaire, B. Lecouteux, D. Schwab, D. Dimitrov, M. S. Hee, M. Ahsan, S. Ahmad, D. Zlatkova, G. Pachov, Z. Xie, P. Nakov, I. Koychev, J. E. Heras Rivera, D. K. Low, W. Yim, J. Ruzevick, D. Child, M. Kurt, Z. Sun, F. Xia, M. Yetisgen, A. Radzhabov, Y. Prokopchuk, V. Kovalev, D. Karpenka, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, M. El Sakka, J. Mothe, A. Băicoianu, C. Florea, M. Ivanovici, Overview of imageclef 2026: Multimodal challenges in medicine, science, agritech, and security, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Springer Lecture Notes in Computer Science LNCS, Jena, Germany, 2026.

[3] H. Damm, T. M. G. Pakull, L. Reinartz, B. Bracke, B. Eryılmaz, P. Nath, R. Brüngel, C. S. Schmidt, H. Schäfer, A. Ben Abacha, A. García Seco de Herrera, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2026 – medical concept detection and caption generation with synthetic data extensions, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026. To appear.

[4] J. Rückert, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, S. Koitka, O. Pelka, A. B. Abacha, A. G. S. de Herrera, H. Müller, P. Horn, F. Nensa, C. M. Friedrich, ROCOv2: Radiology objects in context version 2, an updated multimodal image dataset, Scientific Data 11 (2024). doi:10.1038/s41597-024-03496-6.

[5] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, V. Goel, Self-critical sequence training for image captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[6] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu, D. Guo, Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL: https://arxiv.org/abs/2402.03300. arXiv:2402.03300.

[7] S.-Y. Liu, X. Dong, X. Lu, S. Diao, P. Belcak, M. Liu, M.-H. Chen, H. Yin, Y.-C. F. Wang, K.-T. Cheng, Y. Choi, J. Kautz, P. Molchanov, GDPO: Group reward-decoupled normalization policy optimization for multi-reward rl optimization, 2026. arXiv:2601.05242.

[8] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert., 2020. URL: http://dblp.uni-trier.de/db/conf/iclr/iclr2020.html#ZhangKWWA20.

[9] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, 2004. URL: https://aclanthology.org/W04-1013/.

[10] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong, A. Tupini, Y. Wang, M. Mazzola, S. Shukla, L. Liden, J. Gao, A. Crabtree, B. Piening, C. Bifulco, M. P. Lungren, T. Naumann, S. Wang, H. Poon, A multimodal biomedical foundation model trained from fifteen million image–text pairs, NEJM AI 2 (2024). URL: https://ai.nejm.org/doi/full/10.1056/AIoa2400640. doi:10.1056/AIoa2400640.

[11] Z. Kraljevic, D. Bean, A. Mascio, L. Roguski, A. Folarin, A. Roberts, R. Bendayan, R. J. B. Dobson, Medcat - medical concept annotation tool, CoRR abs/1912.10166 (2019). URL: http://arxiv.org/abs/1912.10166. arXiv:1912.10166.

[12] U.S. National Library of Medicine, Unified medical language system, https://www.nlm.nih.gov/research/umls/index.html, 2026. Accessed 2026-05-18.
[13] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, 2021.
[14] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, 2022. URL: https://doi.org/10.1145/3458754. doi:10.1145/3458754.
[15] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.
[16] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: International Conference on Machine Learning, 2015.
[17] ImageCLEFmedical Caption Organizers, ImageCLEFmedical Caption 2026 task website, https://www.imageclef.org/2026/medical/caption, 2026. Accessed 2026-05-18.
[18] ImageCLEFmedical Caption Organizers, Imageclefmedical caption 2025 overview, https://ceur-ws.org/Vol-4038/paper_170.pdf, 2025.
[19] Y. Lee, H. J. Kim, H. Shin, C. Lim, Ai stat lab: A modular framework for clinically accurate medical image captioning using vision-language models, in: CEUR Workshop Proceedings, volume 4038, CEUR-WS, 2025, pp. 2511–2523.
[20] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, 2021.
[21] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, J. Lin, Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, arXiv preprint arXiv:2409.12191 (2024).
[22] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, J. Lin, Qwen2.5-vl technical report, arXiv preprint arXiv:2502.13923 (2025).
[23] Qwen Team, Qwen3-VL model collection, https://huggingface.co/collections/Qwen/qwen3-vl, 2026. Hugging Face collection, accessed 2026-05-19.
[24] A. Sellergren, S. Kazemzadeh, T. Jaroensri, A. Kiraly, M. Traverse, T. Kohlberger, S. Xu, F. Jamil, C. Hughes, C. Lau, J. Chen, F. Mahvar, L. Yatziv, T. Chen, B. Sterling, S. A. Baby, S. M. Baby, J. Lai, S. Schmidgall, L. Yang, K. Chen, P. Bjornsson, S. Reddy, R. Brush, K. Philbrick, M. Asiedu, I. Mezerreg, H. Hu, H. Yang, R. Tiwari, S. Jansen, P. Singh, Y. Liu, S. Azizi, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Riviere, L. Rouillard, T. Mesnard, G. Cideron, J. bastien Grill, S. Ramos, E. Yvinec, M. Casbon, E. Buchatskaya, J.-B. Alayrac, D. Lepikhin, V. Feinberg, S. Borgeaud, A. Andreev, C. Hardin, R. Dadashi, L. Hussenot, A. Joulin, O. Bachem, Y. Matias, K. Chou, A. Hassidim, K. Goel, C. Farabet, J. Barral, T. Warkentin, J. Shlens, D. Fleet, V. Cotruta, O. Sanseviero, G. Martins, P. Kirk, A. Rao, S. Shetty, D. F. Steiner, C. Kirmizibayrak, R. Pilgrim, D. Golden, L. Yang, Medgemma technical report, 2026. URL: https://arxiv.org/abs/2507.05201. arXiv:2507.05201.
[25] OpenMed, OpenMed/Qwen3.5-2B-MedVL, https://huggingface.co/OpenMed/Qwen3.5-2B-MedVL, 2026. Hugging Face model card, accessed 2026-05-19.
[26] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, R. Lowe, Training language models to follow instructions with human feedback, 2022. URL: http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.
[27] ImageCLEFmedical Caption Organizers, Clef caption evaluation repository 2026, https://github.com/taubsity/clef-caption-evaluation-2026, 2026.
[28] T. Sellam, D. Das, A. P. Parikh, BLEURT: Learning robust metrics for text generation, in: Pro-

ceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020. URL: https://aclanthology.org/2020.acl-main.704/.

[29] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240.

[30] Y. Zha, Y. Yang, R. Li, Z. Hu, AlignScore: Evaluating factual consistency with a unified alignment function, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023. URL: https://aclanthology.org/2023.acl-long.634/.

[31] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, J. Zhou, Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, arXiv preprint arXiv:2308.12966 (2023).

[32] Qwen Team, Qwen3 technical report, 2025. URL: https://arxiv.org/abs/2505.09388. arXiv:2505.09388.

[33] J.-B. Delbrouck, RadGraph-XL: A Large-Scale Expert-Annotated Dataset for Entity and Relation Extraction from Radiology Reports, PhysioNet (2025). URL: https://doi.org/10.13026/j8e7-pr22. doi:10.13026/j8e7-pr22, version 1.0.0.

Appendix

A.1. Training and Inference Hyperparameters

Table A.1 summarizes the main hyperparameters used for supervised fine-tuning, post-SFT GDPO, rollout generation, final inference, and the training environment.

Table A.1: Main training and inference hyperparameters used in the reported runs.

Stage	Hyperparameter	Value
SFT	Optimization	Full-parameter fine-tuning
	Epochs	5
	Optimizer	AdamW (fused)
	Precision	bfloat16
	Learning rate	2×10^{-5}
	Scheduler / warmup	Cosine schedule with 5% warmup
	Effective batch size	128
	Maximum sequence length	2,048
	Maximum image size	802,816 pixels
GDPO	Optimizer	AdamW
	Learning rate	5×10^{-6}
	Maximum steps	3,000
	Warmup ratio	0.01
	Gradient clipping	1.0
	KL coefficient	$\beta = 0.001$
	Clipping parameter	$\epsilon = 0.2$
	Group size	16
	Effective policy batch size	16
	Maximum completion length	100
	Save interval	250 steps
	Rollout sampling	Temperature 1.2, top-p 0.9
Inference	Checkpoint screening	Greedy decoding, beam size 1
	Submission decoding	Beam search, beam size 4
	Maximum new tokens	100
	Length penalty	1.0
	Temperature	0
	Disabled constraints	No minimum-length constraint, no no-repeat n-gram blocking, no top-p or top-k sampling
Environment	Random seed	42
	Hardware	NVIDIA A100 80GB GPU

A.2. Internal Validation Results

Table A.2 reports representative internal validation results used for checkpoint, decoding, and candidate-pool selection. Relevance is computed as the average of the four relevance metrics, factuality as the average of UMLS Concept F1 and AlignScore, and overall as the average of relevance and factuality.

Table A.2: Representative internal validation results used for checkpoint and candidate selection.

Track	Backbone	Setting	Candidate	Rel.	Fact.	Overall
Standard	Qwen3-VL-4B	fixed GDPO	step 2750	0.5663	0.1545	0.3604
	MedGemma-4B	fixed GDPO	step 1500	0.5715	0.1582	0.3649
			beam-4 #1	0.5712	0.1612	0.3662
	Qwen-MedVL-2B	fixed GDPO	step 2750	0.5770	0.1550	0.3660
			beam-4 #1	0.5771	0.1598	0.3684
		piecewise GDPO	step 2750	0.5776	0.1544	0.3660
			beam-4 #1	0.5773	0.1584	0.3678
		dynamic GDPO	step 3000	0.5750	0.1488	0.3619
Synthetical	Qwen3-VL-4B	fixed GDPO	step 2000	0.6905	0.4053	0.5479
	MedGemma-4B	fixed GDPO	step 2000	0.7071	0.4309	0.5690
			beam-4 #1	0.7030	0.4346	0.5688
	Qwen-MedVL-2B	fixed GDPO	step 2750	0.7129	0.4347	0.5738
		piecewise GDPO	step 2750	0.7108	0.4373	0.5740
		dynamic GDPO	step 3000	0.7119	0.4389	0.5738
			beam-4 #1	0.7135	0.4460	0.5797