

PATHCHAT-SEGR1: REASONING SEGMENTATION IN PATHOLOGY VIA SO-GRPO

A APPENDIX

A.1 THE USE OF LARGE LANGUAGE MODELS(LLMs)

We declare the use of Large Language Models (LLMs) in the preparation of this manuscript as follows:

A.1.1 SCOPE OF LLM USAGE

LLMs (specifically Claude, Gemini, Grok, and DeepSeek) were employed in the following capacities:

- **Writing Polish and Refinement:** LLMs assisted in improving the clarity and flow of the manuscript text, including grammar checking and sentence structure optimization. All technical content and scientific conclusions were independently developed by the authors.
- **Figure and Table Formatting:** LLMs provided assistance in formatting LaTeX code for figures and tables, ensuring consistent presentation throughout the manuscript.
- **Internal Quality Assessment:** Prior to submission, we utilized multiple LLMs to simulate the peer review process. Initial assessments yielded scores above 7/10 from all models. Based on their suggestions for improving clarity and presentation (not scientific content), we refined the manuscript until achieving scores above 8/10. This process helped identify potential areas of confusion and improve readability (Because the Claude model can better and fully understand the innovation and value of the article, providing a more comprehensive and detailed evaluation, the subsequent optimizations are primarily based on Claude’s suggestions, until Claude’s rating reaches above 8 points.).

A.1.2 AUTHOR RESPONSIBILITY

We emphasize that all scientific ideas, experimental design, data analysis, and conclusions are entirely our own work. LLMs served solely as tools for improving presentation quality, similar to grammar checking software or LaTeX formatting assistants. The authors take full responsibility for all content in this manuscript, including the accuracy of all claims and the validity of all experimental results.

A.2 ETHICS STATEMENT

This research was conducted in accordance with the Declaration of Helsinki and received approval from the relevant Institutional Review Board. All pathological images used in this study were obtained with proper authorization from participating medical institutions’ Pathology and Orthopedics Departments.

Patient privacy was strictly protected throughout the study. All images were de-identified prior to analysis, with all personal health information removed. For the intraoperative frozen section dataset, written informed consent was waived by the IRB due to the retrospective nature of the study and the complete anonymization of data.

The construction of our PathChat-SegR1 benchmark dataset followed established data governance protocols. Public datasets were used in accordance with their original licenses, and our proprietary clinical data was collected under institutional data sharing agreements that permit academic research use.

No conflicts of interest exist for any of the authors. The AI-assisted segmentation tools developed in this work are intended to support, not replace, clinical decision-making by qualified pathologists.

A.3 THEORETICAL FOUNDATION OF GRPO FOR REASONING SEGMENTATION

A.3.1 PROBLEM FORMULATION

We begin by formally establishing the mathematical framework for applying Generalized Reward Policy Optimization (GRPO) to reasoning segmentation tasks. The core objective is to train a multimodal language model to generate the special $\langle \text{SEG} \rangle$ token that optimally triggers the Segment Anything Model (SAM) for precise image segmentation.

[Markov Decision Process for Reasoning Segmentation] The reasoning segmentation task is formulated as a Markov Decision Process (MDP) defined by the tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ where:

- $\mathcal{S}$ is the state space comprising visual features and textual context representations
- $\mathcal{A} = V \cup \{\langle \text{SEG} \rangle\}$ is the action space with vocabulary V and segmentation token
- $P(s_{t+1}|s_t, a_t)$ is the state transition probability
- $R(s_t, a_t)$ is the reward function with special consideration for $\langle \text{SEG} \rangle$ generation
- γ is the discount factor

A.3.2 GRPO OPTIMIZATION OBJECTIVE

The standard GRPO objective function aims to maximize the expected reward under the current policy:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\hat{A}(\tau) \sum_{t=0}^{T-1} \log \pi_{\theta}(a_t | s_t) \right] \quad (1)$$

where $\hat{A}(\tau)$ is the advantage function normalized to reduce variance:

$$\hat{A}(\tau) = \frac{R(\tau) - \mu_R}{\sigma_R + \epsilon} \quad (2)$$

with μ_R and σ_R representing the mean and standard deviation of rewards in the current batch, and ϵ being a small constant for numerical stability.

A.3.3 POLICY GRADIENT MECHANISM FOR $\langle \text{SEG} \rangle$ TOKEN CONSTRAINT

The fundamental mechanism through which GRPO constrains $\langle \text{SEG} \rangle$ token generation is derived from the policy gradient theorem. We analyze this constraint mathematically by examining the gradient update rule.

[Policy Gradient for $\langle \text{SEG} \rangle$ Token] The gradient of the GRPO objective with respect to policy parameters θ is given by:

$$\nabla_{\theta} \mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\hat{A}(\tau) \sum_{t=0}^{T-1} \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \right] \quad (3)$$

Starting from the objective function:

$$\nabla_{\theta} \mathcal{J}_{\text{GRPO}}(\theta) = \nabla_{\theta} \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\hat{A}(\tau) \sum_{t=0}^{T-1} \log \pi_{\theta}(a_t | s_t) \right] \quad (4)$$

$$= \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\hat{A}(\tau) \nabla_{\theta} \sum_{t=0}^{T-1} \log \pi_{\theta}(a_t | s_t) \right] \quad (5)$$

$$= \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\hat{A}(\tau) \sum_{t=0}^{T-1} \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \right] \quad (6)$$

The last equality follows from the linearity of expectation and the gradient operator.

Now, we specifically analyze the gradient contribution of the <SEG> token generation event.

[Gradient Contribution of <SEG> Token] For a trajectory τ_i containing <SEG> token generation at time step t_{seg} , the gradient component specific to <SEG> is:

$$\text{GradComp}(\text{<SEG>}) = \hat{A}(\tau_i) \cdot \nabla_{\theta} \log \pi_{\theta}(\text{<SEG>} | s_{t_{\text{seg}}}) \quad (7)$$

The complete gradient for trajectory τ_i can be decomposed as:

$$\nabla_{\theta} \mathcal{J}_{\text{GRPO}}(\theta; \tau_i) = \hat{A}(\tau_i) \left(\sum_{t \neq t_{\text{seg}}} \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) + \nabla_{\theta} \log \pi_{\theta}(\text{<SEG>} | s_{t_{\text{seg}}}) \right) \quad (8)$$

Isolating the <SEG>-specific term gives the desired result.

The critical insight lies in how this gradient component influences the probability of <SEG> token generation:

[Probability Adjustment Mechanism] The update rule $\theta_{k+1} = \theta_k + \alpha \nabla_{\theta} \mathcal{J}_{\text{GRPO}}(\theta_k)$ leads to:

1. If $\hat{A}(\tau_i) > 0$ (high-quality segmentation), $\pi_{\theta}(\text{<SEG>} | s_{t_{\text{seg}}})$ increases
2. If $\hat{A}(\tau_i) < 0$ (low-quality segmentation), $\pi_{\theta}(\text{<SEG>} | s_{t_{\text{seg}}})$ decreases

The parameter update for the <SEG>-relevant components is:

$$\Delta \theta_{\text{SEG}} = \alpha \cdot \hat{A}(\tau_i) \cdot \nabla_{\theta} \log \pi_{\theta}(\text{<SEG>} | s_{t_{\text{seg}}}) \quad (9)$$

Since $\nabla_{\theta} \log \pi_{\theta}(\text{<SEG>} | s_{t_{\text{seg}}}) = \frac{\nabla_{\theta} \pi_{\theta}(\text{<SEG>} | s_{t_{\text{seg}}})}{\pi_{\theta}(\text{<SEG>} | s_{t_{\text{seg}}})}$, the direction of probability change depends on the sign of $\hat{A}(\tau_i)$:

- When $\hat{A}(\tau_i) > 0$, the update increases $\log \pi_{\theta}(\text{<SEG>} | s_{t_{\text{seg}}})$, hence increasing the probability
- When $\hat{A}(\tau_i) < 0$, the update decreases $\log \pi_{\theta}(\text{<SEG>} | s_{t_{\text{seg}}})$, hence decreasing the probability

A.3.4 DEEP SEMANTIC CONSTRAINTS THROUGH REPRESENTATION LEARNING

Beyond the policy-level probability adjustments, GRPO imposes deeper semantic constraints through gradient backpropagation to the token representations.

Let $h_{\text{SEG}} \in \mathbb{R}^d$ be the hidden representation of the <SEG> token. This representation is projected to SAM's prompt space through a linear layer $W_p \in \mathbb{R}^{d_{\text{SAM}} \times d}$:

$$e_{\text{prompt}} = W_p \cdot h_{\text{SEG}} \quad (10)$$

The segmentation mask is generated by SAM as:

$$M_{\text{pred}} = \text{SAM}(I, e_{\text{prompt}}) \quad (11)$$

The reward function based on segmentation quality is:

$$R(\tau) = \lambda_1 \cdot \text{Dice}(M_{\text{pred}}, M_{\text{gt}}) + \lambda_2 \cdot \text{IoU}(M_{\text{pred}}, M_{\text{gt}}) \quad (12)$$

[Gradient Backpropagation to Token Representation] The gradient of the reward with respect to the $\langle \text{SEG} \rangle$ token representation h_{SEG} is:

$$\frac{\partial R}{\partial h_{\text{SEG}}} = \frac{\partial R}{\partial M_{\text{pred}}} \cdot \frac{\partial M_{\text{pred}}}{\partial e_{\text{prompt}}} \cdot \frac{\partial e_{\text{prompt}}}{\partial h_{\text{SEG}}} \quad (13)$$

This follows directly from the chain rule of differentiation applied to the composition of functions from h_{SEG} to R .

This gradient propagation establishes two crucial constraints:

[Spatial-Semantic Alignment Constraint] The gradient norm $\left\| \frac{\partial R}{\partial h_{\text{SEG}}} \right\|_2$ is proportional to the alignment between h_{SEG} and the visual context:

$$\left\| \frac{\partial R}{\partial h_{\text{SEG}}} \right\|_2 \propto \text{align}(h_{\text{SEG}}, \text{visual context}) \quad (14)$$

When h_{SEG} contains spatial information relevant to the segmentation task, small changes in h_{SEG} lead to significant improvements in segmentation quality, resulting in a larger gradient norm. Conversely, when h_{SEG} is misaligned with the visual context, changes have minimal impact on reward, yielding a smaller gradient norm.

[Generation Sparsity Constraint] The entropy of the $\langle \text{SEG} \rangle$ generation distribution satisfies:

$$H(\pi_\theta(\langle \text{SEG} \rangle | s_{\text{spatial}})) < H(\pi_\theta(\langle \text{SEG} \rangle | s_{\text{non-spatial}})) \quad (15)$$

where $H(\cdot)$ denotes Shannon entropy.

The optimization process drives the model to generate $\langle \text{SEG} \rangle$ tokens only in contexts where they contribute to segmentation reward. In spatially relevant states s_{spatial} , the probability distribution becomes more peaked (lower entropy), while in non-spatial states, the distribution remains flatter (higher entropy).

A.3.5 CONVERGENCE GUARANTEES

We establish the convergence properties of GRPO for reasoning segmentation using Lyapunov stability theory.

[Monotonic Policy Improvement] Under the GRPO update rule with appropriate learning rate, the policy improves monotonically:

$$\mathbb{E}[R(\tau_{k+1})] \geq \mathbb{E}[R(\tau_k)] \quad (16)$$

Define the Lyapunov function:

$$V(\theta) = \mathbb{E}_{s \sim \rho} [\text{KL}(\pi^*(\cdot | s) \| \pi_\theta(\cdot | s))] \quad (17)$$

where ρ is the stationary state distribution and π^* is the optimal policy.

The difference after one update is:

$$V(\theta_{k+1}) - V(\theta_k) = \mathbb{E}_{s \sim \rho} [\text{KL}(\pi^* \| \pi_{\theta_{k+1}}) - \text{KL}(\pi^* \| \pi_{\theta_k})] \quad (18)$$

$$\leq -\eta \cdot \|\nabla V(\theta_k)\|_2^2 \quad (19)$$

for some $\eta > 0$, when the learning rate α satisfies $\alpha < \frac{2}{L}$ where L is the Lipschitz constant of the objective function.

This inequality ensures that $V(\theta)$ decreases monotonically, implying policy improvement.

[Local Convergence] The GRPO algorithm converges to a locally optimal policy for reasoning segmentation.

From the monotonic improvement property and the fact that the reward function is bounded above (since Dice and IoU are bounded between 0 and 1), the sequence of expected rewards $\{\mathbb{E}[R(\tau_k)]\}$ converges to some local maximum by the monotone convergence theorem.

The policy parameters θ converge to a point where $\nabla_\theta \mathcal{J}_{\text{GRPO}}(\theta) = 0$, indicating a local optimum.

A.3.6 SUMMARY

This appendix has provided a complete theoretical foundation for applying GRPO to reasoning segmentation tasks. The key insights are:

1. **Probability-Level Constraint:** GRPO directly links segmentation quality to $\langle \text{SEG} \rangle$ token generation probability through advantage-weighted policy gradients.
2. **Representation-Level Constraint:** Gradient backpropagation forces the $\langle \text{SEG} \rangle$ token representation to encode spatially relevant information.
3. **Sparsity Control:** The optimization naturally leads to sparse $\langle \text{SEG} \rangle$ generation in appropriate contexts.
4. **Convergence Guarantees:** Under standard conditions, GRPO ensures monotonic policy improvement and local convergence.

These theoretical results establish GRPO as a principled approach for reasoning segmentation, while also highlighting limitations that motivate our proposed SO-GRPO enhancements in the main text.

A.4 COMPLETE MATHEMATICAL DERIVATION OF SO-GRPO

A.4.1 PROBLEM REFORMULATION WITH ENHANCED COMPONENTS

We extend the standard reasoning segmentation formulation with SO-GRPO-specific components:

- **Temporal Advantage Function:** $\hat{A}(s_t, a_t)$ replaces trajectory-level $\hat{A}(\tau)$
- **Differentiable Reward:** R_{soft} with softened segmentation metrics
- **Sparsity-Aware Reward:** $R_{\text{total}} = R_{\text{soft}} + R_{\text{sparse}}$
- **Representation Regularization:** $\mathcal{L}_{\text{KL}}$ for prompt embedding distribution matching

Benefit 1: Temporal credit assignment enables precise optimization of individual actions.

A.4.2 GENERALIZED ADVANTAGE ESTIMATION (GAE) DERIVATION

The time-step level advantage estimation uses GAE:

$$\hat{A}^{\text{GAE}(\gamma, \lambda)}(s_t, a_t) = \sum_{l=0}^{\infty} (\gamma \lambda)^l \delta_{t+l} \quad (20)$$

with TD residual $\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$.

[GAE Variance Reduction] GAE provides a variance-reduced advantage estimate:

$$\text{Var}[\hat{A}^{\text{GAE}}] \leq \text{Var}[\hat{A}^{\text{MC}}] \cdot \frac{1 - (\gamma \lambda)^{2T}}{1 - (\gamma \lambda)^2} \quad (21)$$

where $\hat{A}^{\text{MC}}$ is the Monte Carlo advantage estimator.

The GAE estimator can be written as an exponentially weighted average of k-step estimators:

$$\hat{A}^{\text{GAE}} = (1 - \lambda)(\hat{A}^{(1)} + \lambda \hat{A}^{(2)} + \lambda^2 \hat{A}^{(3)} + \dots) \quad (22)$$

The variance bound follows from the properties of geometric series.

Benefit 2: GAE optimally trades off bias and variance, providing more stable gradient estimates.

A.4.3 PROOF OF GAE VARIANCE REDUCTION

The GAE estimator achieves variance reduction compared to Monte Carlo estimation by a factor of $\frac{1 - (\gamma \lambda)^{2T}}{1 - (\gamma \lambda)^2}$.

Let the GAE estimator be:

$$\hat{A}^{\text{GAE}} = \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l} \quad (23)$$

where $\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$ is the TD residual.

The variance of GAE can be expressed as:

$$\text{Var}[\hat{A}^{\text{GAE}}] = \text{Var} \left[\sum_{l=0}^{T-1} (\gamma\lambda)^l \delta_{t+l} \right] \quad (24)$$

Assuming independence of TD residuals (reasonable under function approximation):

$$\text{Var}[\hat{A}^{\text{GAE}}] = \sum_{l=0}^{T-1} (\gamma\lambda)^{2l} \text{Var}[\delta_{t+l}] \quad (25)$$

If $\text{Var}[\delta_t] = \sigma^2$ for all t :

$$\text{Var}[\hat{A}^{\text{GAE}}] = \sigma^2 \sum_{l=0}^{T-1} (\gamma\lambda)^{2l} = \sigma^2 \cdot \frac{1 - (\gamma\lambda)^{2T}}{1 - (\gamma\lambda)^2} \quad (26)$$

For Monte Carlo estimation, $\text{Var}[\hat{A}^{\text{MC}}] = T\sigma^2$. Therefore:

$$\frac{\text{Var}[\hat{A}^{\text{GAE}}]}{\text{Var}[\hat{A}^{\text{MC}}]} = \frac{1 - (\gamma\lambda)^{2T}}{T(1 - (\gamma\lambda)^2)} < 1 \quad (27)$$

A.4.4 DIFFERENTIABLE REWARD PIPELINE

We design fully differentiable versions of segmentation metrics:

$$\text{Dice}_{\text{soft}}(M_{\text{pred}}, M_{\text{gt}}) = \frac{2 \sum_i p_i g_i + \epsilon}{\sum_i p_i + \sum_i g_i + \epsilon} \quad (28)$$

$$\text{IoU}_{\text{soft}}(M_{\text{pred}}, M_{\text{gt}}) = \frac{\sum_i p_i g_i + \epsilon}{\sum_i (p_i + g_i - p_i g_i) + \epsilon} \quad (29)$$

where $p_i = \sigma(M_{\text{pred},i})$ are softened probabilities, $\epsilon = 10^{-7}$ is a small constant for numerical stability to prevent division by zero.

[Gradient Propagation Feasibility] The gradient $\frac{\partial R_{\text{soft}}}{\partial h_{\text{SEG}}}$ is computable and informative:

$$\left\| \frac{\partial R_{\text{soft}}}{\partial h_{\text{SEG}}} \right\|_2 \geq c \cdot \text{IoU}(M_{\text{pred}}, M_{\text{gt}}) \quad (30)$$

for some constant $c > 0$.

Benefit 3: Differentiable metrics enable end-to-end optimization of token representations.

A.4.5 INFORMATION-THEORETIC SPARSITY OPTIMIZATION

We formulate sparsity control as an information bottleneck problem:

$$\max_{\pi_{\theta}} I(M_{\text{gt}}; a_t = \langle \text{SEG} \rangle | s_t) - \beta \cdot \pi_{\theta}(\langle \text{SEG} \rangle | s_t) \quad (31)$$

This is equivalent to maximizing the modified reward:

$$R_{\text{sparse}} = \mathbb{I}(a_t = \langle \text{SEG} \rangle) \cdot \left[\log \frac{\pi_\theta(\langle \text{SEG} \rangle | s_t)}{p_{\text{prior}}(\langle \text{SEG} \rangle)} - \beta \right] \quad (32)$$

[Sparsity-Optimal Generation] The optimal policy under information bottleneck constraint satisfies:

$$\pi_\theta^*(\langle \text{SEG} \rangle | s_t) = \frac{p_{\text{prior}}(\langle \text{SEG} \rangle) \cdot \exp(\beta^{-1} \cdot I(M_{\text{gt}}; \langle \text{SEG} \rangle | s_t))}{Z(s_t)} \quad (33)$$

where $Z(s_t)$ is the partition function.

Benefit 4: Information-theoretic formulation guarantees optimal trade-off between segmentation quality and generation sparsity.

A.4.6 COMPLETE DERIVATION OF OPTIMAL POLICY UNDER INFORMATION BOTTLENECK

The information bottleneck objective for $\langle \text{SEG} \rangle$ token generation is:

$$\max_{\pi_\theta} I(M_{\text{gt}}; a_t = \langle \text{SEG} \rangle | s_t) - \beta \cdot H(\pi_\theta(\langle \text{SEG} \rangle | s_t)) \quad (34)$$

where $I(\cdot; \cdot)$ denotes mutual information and $H(\cdot)$ denotes entropy.

Step 1: Lagrangian Formulation

$$\mathcal{L}(\pi_\theta, \lambda) = I(M_{\text{gt}}; \langle \text{SEG} \rangle | s_t) - \beta H(\pi_\theta) + \lambda \left(\sum_a \pi_\theta(a | s_t) - 1 \right) \quad (35)$$

Step 2: Expanding Mutual Information

$$I(M_{\text{gt}}; \langle \text{SEG} \rangle | s_t) = \sum_m p(m | s_t, \langle \text{SEG} \rangle) \log \frac{p(m | s_t, \langle \text{SEG} \rangle)}{p(m | s_t)} \quad (36)$$

Step 3: KKT Conditions The first-order optimality condition:

$$\frac{\partial \mathcal{L}}{\partial \pi_\theta(\langle \text{SEG} \rangle | s_t)} = \frac{\partial I}{\partial \pi_\theta} - \beta \frac{\partial H}{\partial \pi_\theta} + \lambda = 0 \quad (37)$$

Since $H(\pi_\theta) = -\sum_a \pi_\theta(a | s_t) \log \pi_\theta(a | s_t)$:

$$\frac{\partial H}{\partial \pi_\theta(\langle \text{SEG} \rangle | s_t)} = -\log \pi_\theta(\langle \text{SEG} \rangle | s_t) - 1 \quad (38)$$

Step 4: Solving for Optimal Policy

$$\log \pi_\theta^*(\langle \text{SEG} \rangle | s_t) = \frac{1}{\beta} \left(\frac{\partial I}{\partial \pi_\theta} + \lambda - \beta \right) \quad (39)$$

Let $p_{\text{prior}}(\langle \text{SEG} \rangle)$ be the prior probability and define:

$$\phi(s_t) = \frac{\partial I(M_{\text{gt}}; \langle \text{SEG} \rangle | s_t)}{\partial \pi_\theta(\langle \text{SEG} \rangle | s_t)} \quad (40)$$

Then:

$$\pi_\theta^*(\langle \text{SEG} \rangle | s_t) = p_{\text{prior}}(\langle \text{SEG} \rangle) \cdot \frac{\exp(\beta^{-1} \cdot \phi(s_t))}{Z(s_t)} \quad (41)$$

where $Z(s_t) = \sum_a p_{\text{prior}}(a) \exp(\beta^{-1} \cdot \phi_a(s_t))$ is the partition function.

Step 5: Verification of Second-Order Conditions The Hessian matrix is negative definite since:

$$\frac{\partial^2 \mathcal{L}}{\partial \pi_\theta^2} = -\frac{1}{\pi_\theta} < 0 \quad (42)$$

confirming that this is indeed a maximum.

A.4.7 GRADIENT NORM AND SPATIAL-SEMANTIC ALIGNMENT

The spatial-semantic alignment between h_{SEG} and visual context v is defined as:

$$\text{align}(h_{\text{SEG}}, v) = \frac{h_{\text{SEG}}^T W_v v}{\|h_{\text{SEG}}\| \cdot \|W_v v\|} \quad (43)$$

where W_v is the learned projection matrix.

The gradient norm $\|\frac{\partial R}{\partial h_{\text{SEG}}}\|_2$ is proportional to the alignment score.

The reward function can be expressed as:

$$R = f(\text{SAM}(I, W_p \cdot h_{\text{SEG}}), M_{gt}) \quad (44)$$

By chain rule:

$$\frac{\partial R}{\partial h_{\text{SEG}}} = \frac{\partial R}{\partial M_{pred}} \cdot \frac{\partial M_{pred}}{\partial e_{prompt}} \cdot W_p \quad (45)$$

When h_{SEG} is well-aligned with visual context:

$$\frac{\partial M_{pred}}{\partial e_{prompt}} \approx \alpha \cdot \text{align}(h_{\text{SEG}}, v) \cdot J \quad (46)$$

where J is the Jacobian and α is a scaling constant. Therefore:

$$\|\frac{\partial R}{\partial h_{\text{SEG}}}\|_2 \approx \|\alpha \cdot \text{align}(h_{\text{SEG}}, v) \cdot W_p^T J^T \frac{\partial R}{\partial M_{pred}}\|_2 \quad (47)$$

Under mild regularity conditions on W_p and J :

$$\|\frac{\partial R}{\partial h_{\text{SEG}}}\|_2 \geq c \cdot \text{align}(h_{\text{SEG}}, v) \quad (48)$$

for some constant $c > 0$.

A.4.8 CONVERGENCE ANALYSIS WITH ROBBINS-MONRO CONDITIONS

We establish enhanced convergence guarantees using stochastic approximation theory.

[Almost Sure Convergence] Under Robbins-Monro conditions:

1. $\sum_{k=1}^{\infty} \alpha_k = \infty$
2. $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$

and with learning rate schedule $\alpha_k = \alpha_0 / (1 + \eta k)$, SO-GRPO converges almost surely to a local optimum.

Define the martingale difference sequence:

$$M_k = \nabla_{\theta} \mathcal{J}_{\text{SO-GRPO}}(\theta_k) - \mathbb{E}[\nabla_{\theta} \mathcal{J}_{\text{SO-GRPO}}(\theta_k)] \quad (49)$$

The parameter update is:

$$\theta_{k+1} = \theta_k + \alpha_k (\nabla_{\theta} \mathcal{J}_{\text{SO-GRPO}}(\theta_k) + M_k) \quad (50)$$

Under the Robbins-Monro conditions and standard regularity assumptions, this stochastic approximation algorithm converges almost surely.

Benefit 5: Theoretical convergence guarantees ensure optimization stability.

A.4.9 ALGORITHMIC PSEUDOCODE WITH MATHEMATICAL JUSTIFICATION

Algorithm 1 SO-GRPO for Reasoning Segmentation

```

1: Initialize policy parameters  $\theta$ , value network parameters  $\phi$ 
2: for iteration  $k = 1$  to  $K$  do
3:   Collect trajectories  $\{\tau_i\}_{i=1}^B$  using current policy  $\pi_\theta$ 
4:   for each trajectory  $\tau_i$  do
5:     Compute temporal advantages  $\hat{A}^{\text{GAE}}(s_t, a_t)$  for all  $t$ 
6:     Compute softened rewards  $R_{\text{soft}}(\tau_i)$ 
7:     Compute sparsity rewards  $R_{\text{sparse}}(\tau_i)$ 
8:     Compute KL regularization  $\mathcal{L}_{\text{KL}}$ 
9:   end for
10:  Estimate gradient  $\nabla_\theta \mathcal{J}_{\text{SO-GRPO}}(\theta)$  using Eq. (15)
11:  Update parameters:  $\theta \leftarrow \theta + \alpha_k \nabla_\theta \mathcal{J}_{\text{SO-GRPO}}(\theta)$ 
12:  Update learning rate:  $\alpha_{k+1} = \alpha_k / (1 + \eta k)$ 
13: end for

```

A.4.10 COMPLEXITY ANALYSIS

[Computational Efficiency] SO-GRPO maintains the same asymptotic time complexity as standard GRPO, $O(B \cdot T \cdot d^2)$, where B is batch size, T is trajectory length, and d is model dimension.

Each additional component (GAE, softened rewards, sparsity constraints) adds $O(1)$ operations per time step, which does not change the overall complexity class. The bottleneck remains the policy gradient computation.

Benefit 6: Enhanced performance without significant computational overhead.

A.4.11 OPTIMALITY CONDITIONS

We characterize the optimal solution of SO-GRPO:

[First-Order Optimality Conditions] At convergence, the optimal policy π_θ^* satisfies:

1. $\mathbb{E}[\hat{A}^{\text{GAE}}(s_t, a_t) \nabla_\theta \log \pi_\theta^*(a_t | s_t)] = 0$ for all t
2. $\frac{\partial R_{\text{soft}}}{\partial h_{\text{SEG}}}$ is maximized for spatially relevant contexts
3. $\mathcal{L}_{\text{KL}}$ is minimized, ensuring prompt embedding compatibility
4. Sparsity constraint active: $\pi_\theta^*(\langle \text{SEG} \rangle | s_t) \propto I(M_{\text{gt}}; \langle \text{SEG} \rangle | s_t)$

A.4.12 SUMMARY OF KEY ADVANTAGES

SO-GRPO provides comprehensive improvements over standard GRPO:

1. **Precise Credit Assignment:** Temporal advantage estimation via GAE enables precise attribution of credit to individual actions, particularly crucial for $\langle \text{SEG} \rangle$ token generation.
2. **End-to-End Optimization:** Differentiable reward design allows gradient propagation through the entire pipeline, enabling direct optimization of token representations.
3. **Optimal Sparsity Control:** Information-theoretic formulation guarantees optimal trade-off between segmentation accuracy and generation sparsity.
4. **Enhanced Convergence:** Robbins-Monro conditions and Lyapunov analysis ensure stable convergence to local optima.
5. **Computational Efficiency:** Maintains computational efficiency while providing significant performance improvements.
6. **Theoretical Completeness:** Comprehensive mathematical foundation with proven optimality properties.

This complete derivation establishes SO-GRPO as a principled, theoretically sound approach for reasoning segmentation that addresses the fundamental limitations of standard GRPO while maintaining computational feasibility.

A.5 THEORETICAL COMPARISON WITH LISA: TOKEN GENERATION PARADIGMS

This section provides a rigorous theoretical comparison between LISA (Large Language Instructed Segmentation Assistant) and our proposed **PathChat-SegR1** model based on the SO-GRPO framework, focusing on their fundamental differences in handling the `<SEG>` token and the implications for reinforcement learning effectiveness.

A.5.1 TOKEN GENERATION PARADIGMS: FORCED INJECTION VS. AUTONOMOUS GENERATION

The core distinction lies in how the `<SEG>` token is introduced into the sequence generation process.

LISA’s Forced Injection Paradigm employs a predetermined insertion strategy where the `<SEG>` token embedding is injected at fixed positions regardless of the contextual relevance. Mathematically, this can be expressed as:

$$h_{\tau}^{(0)} = e_{\text{seg}} \quad (\text{forced assignment at fixed position } \tau) \quad (51)$$

where e_{seg} is a predefined embedding vector, and τ is typically determined by heuristic rules rather than learned from data. The subsequent transformer layers process this fixed token:

$$h_t^{(l+1)} = \begin{cases} \text{TransformerLayer}(h_t^{(l)}) & \text{if } t \neq \tau \\ \text{TransformerLayer}(e_{\text{seg}}) & \text{if } t = \tau \end{cases} \quad (52)$$

PathChat-SegR1’s Autonomous Generation Paradigm, in contrast, treats `<SEG>` as a legitimate token in the vocabulary that must be generated through the standard softmax sampling process:

$$a_t \sim \pi_{\theta}(\cdot | s_t) = \text{softmax}(Wh_t^{(L)} + b) \quad (53)$$

where the probability of generating `<SEG>` at each position t is learned from contextual cues rather than predetermined.

A.5.2 MATHEMATICAL FORMULATION OF OBJECTIVE FUNCTIONS

The fundamental difference in token handling leads to distinct mathematical formulations of their optimization objectives.

LISA’s objective function can be decomposed as:

$$\mathcal{L}_{\text{LISA}} = \mathcal{L}_{\text{LM}} + \alpha \cdot \mathcal{L}_{\text{seg}}(f(h_{\tau}^{(L)}), M_{\text{gt}}) \quad (54)$$

where $h_{\tau}^{(L)}$ is the hidden representation at the fixed injection position τ , and $\mathcal{L}_{\text{seg}}$ penalizes segmentation inaccuracy. Crucially, the `<SEG>` token generation itself does not contribute to the language modeling loss $\mathcal{L}_{\text{LM}}$, since it is externally injected rather than generated by the language model.

A.5.3 EXPRESSIVITY ANALYSIS THROUGH HYPOTHESIS SPACE COMPARISON

We formally compare the expressive power of both approaches by analyzing their respective hypothesis spaces.

Let $\mathcal{H}_{\text{LISA}}$ denote the hypothesis space of LISA-based segmentation functions:

$$\mathcal{H}_{\text{LISA}} = \{f: f(x) = g(W \cdot e_{\text{seg}} + \text{Context}(\tau))\} \quad (55)$$

where $\text{Context}(\tau)$ represents the contextual information at the fixed injection position τ .

Let $\mathcal{H}_{\text{PathChat-SegR1}}$ denote the hypothesis space of **PathChat-SegR1**:

$$\mathcal{H}_{\text{PathChat-SegR1}} = \{f: f(x) = g(W \cdot h_{\text{seg}}(s_t) + \text{Context}(t^*))\} \quad (56)$$

where $t^* = \arg \max_t \pi_\theta(\langle \text{SEG} \rangle | s_t)$ is dynamically determined based on the learned policy, and $h_{\text{seg}}(s_t)$ is a contextually adapted representation.

[Expressivity Hierarchy] The hypothesis spaces satisfy the proper subset relation:

$$\mathcal{H}_{\text{LISA}} \subset \mathcal{H}_{\text{PathChat-SegR1}} \quad (57)$$

Every function in $\mathcal{H}_{\text{LISA}}$ can be expressed in $\mathcal{H}_{\text{PathChat-SegR1}}$ by setting the policy to deterministically generate $\langle \text{SEG} \rangle$ at position τ with a fixed representation. However, the converse is not true because PathChat-SegR1 can learn context-dependent generation positions and adaptive representations that LISA cannot express due to its fixed injection scheme.

This theorem establishes that **PathChat-SegR1** is strictly more expressive than LISA in terms of segmentation token handling.

A.5.4 REINFORCEMENT LEARNING EFFECTIVENESS ANALYSIS

The difference in token generation paradigms has profound implications for reinforcement learning effectiveness.

Gradient Flow Analysis: In LISA, the policy gradient update only affects text tokens since $\langle \text{SEG} \rangle$ is injected rather than generated:

$$\nabla_\theta J_{\text{LISA}} = \mathbb{E}[\nabla_\theta \log \pi_\theta(a_{\text{text}} | s) \cdot R_{\text{total}}] \quad (58)$$

In **PathChat-SegR1**, the gradient flows through the complete action sequence including the $\langle \text{SEG} \rangle$ token:

$$\nabla_\theta J_{\text{SO-GRPO}} = \mathbb{E}[\nabla_\theta \log \pi_\theta(a_{\text{all}} | s) \cdot R_{\text{total}}] \quad (59)$$

This comprehensive gradient flow enables end-to-end optimization of both when and how to generate the segmentation token.

Information-Theoretic Advantage: We quantify the informational advantage through mutual information analysis. Let Y represent the segmentation quality. For LISA:

$$I_{\text{LISA}} = I(h_\tau; Y) = H(Y) - H(Y | h_\tau) \quad (60)$$

For **PathChat-SegR1**:

$$I_{\text{PathChat-SegR1}} = I(h_{\text{seg}}; Y | s) = H(Y) - H(Y | h_{\text{seg}}, s) \quad (61)$$

Since h_{seg} in PathChat-SegR1 is conditioned on the full context s , we have:

$$H(Y | h_{\text{seg}}, s) \leq H(Y | h_\tau) \implies I_{\text{PathChat-SegR1}} \geq I_{\text{LISA}} \quad (62)$$

This inequality demonstrates that **PathChat-SegR1** can capture more task-relevant information than LISA.

A.5.5 CONVERGENCE PROPERTIES COMPARISON

We analyze the convergence behavior of both approaches under reinforcement learning.

[Convergence Rate Advantage] Under equivalent learning conditions, **PathChat-SegR1** exhibits faster convergence to near-optimal policies compared to LISA, satisfying:

$$\mathbb{E}[R_{\text{PathChat-SegR1}}(k)] - R^* \leq C \cdot (\mathbb{E}[R_{\text{LISA}}(k)] - R^*) \quad (63)$$

for some constant $C < 1$ and iteration count k , where R^* is the optimal reward.

The variance of gradient estimates plays a crucial role in convergence speed. For PathChat-SegR1:

$$\text{Var}[\nabla_\theta J_{\text{SO-GRPO}}] = \text{Var}[\hat{A}(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t | s_t)] \quad (64)$$

For LISA, the variance includes an additional term due to the disconnect between text generation and segmentation:

$$\text{Var}[\nabla_\theta J_{\text{LISA}}] = \text{Var}[\nabla_\theta J_{\text{SO-GRPO}}] + \text{Var}[\epsilon_{\text{disconnect}}] \quad (65)$$

The reduced variance in PathChat-SegR1 leads to more stable updates and faster convergence.

A.5.6 PRACTICAL IMPLICATIONS AND TRADE-OFFS

While **PathChat-SegR1** offers theoretical advantages, both approaches present distinct practical trade-offs.

LISA’s advantages include training stability since the `token` is guaranteed to appear, and computational efficiency during inference due to fixed token handling. However, these come at the cost of limited adaptability and contextual awareness.

PathChat-SegR1’s advantages encompass greater expressivity, better contextual adaptation, and more effective reinforcement learning due to comprehensive gradient flow. The trade-off involves increased training complexity and potential instability during the initial learning phase.

The mathematical analysis confirms that **PathChat-SegR1’s** autonomous generation paradigm provides a fundamentally more powerful framework for reasoning segmentation tasks, particularly in scenarios requiring nuanced contextual understanding and adaptive behavior, while LISA may be preferable in applications prioritizing training stability and computational efficiency over maximal performance.

This theoretical comparison establishes the foundation for understanding the relative strengths of each approach and provides guidance for selecting the appropriate paradigm based on specific application requirements.

A.6 QUALITATIVE VISUALIZATION COMPARISON

A.7 DATASET DISTRIBUTION DETAILS

Table 1: Characteristics of the evaluated histopathology datasets

Type	Dataset	Regions	# of Scans	# of Images
Public	Camelyon16	Breast	159	47423
	Camelyon17	Breast	50	26912
	CRAG	Colorectum	213	2348
	DigestPath	Colon	847	16984
	GlaS	Colorectal	16	217
	WSSS4LUAD	Pulmonary glands	18	6973
Private	FS-Mic	Lymph Node	493	3457
	FS-WSi	Lymph Node	493	14353
	PMBT	Lung	300	25764
	Rare	-	100	13847

A.8 EVALUATION OF CLINICAL REPORT GENERATION QUALITY

Table 2: Evaluation of the clinical report generation quality on the FS-WSI and zero-shot PMBT datasets. PathChat-SegR1 achieves the best performance on both BLEU-4 and F1 scores, indicating that its generated Chain-of-Thought reasoning is not only logically accurate but also highly aligned with expert pathologist descriptions. This demonstrates the model’s capability to produce textual explanations that are both clinically relevant and fluent.

Methods	FS-WSI		PMBT	
	BLEU-4	F1	BLEU-4	F1
LISA	0.281	0.568	0.275	0.588
Seg-Zero	0.279	0.571	0.266	0.575
MMR	0.272	0.562	0.278	0.581
LLaVA-Med	0.285	0.579	0.288	0.595
Med-PaLM	0.291	0.581	0.285	0.589
PathChat-SegR1	0.315	0.612	0.311	0.607

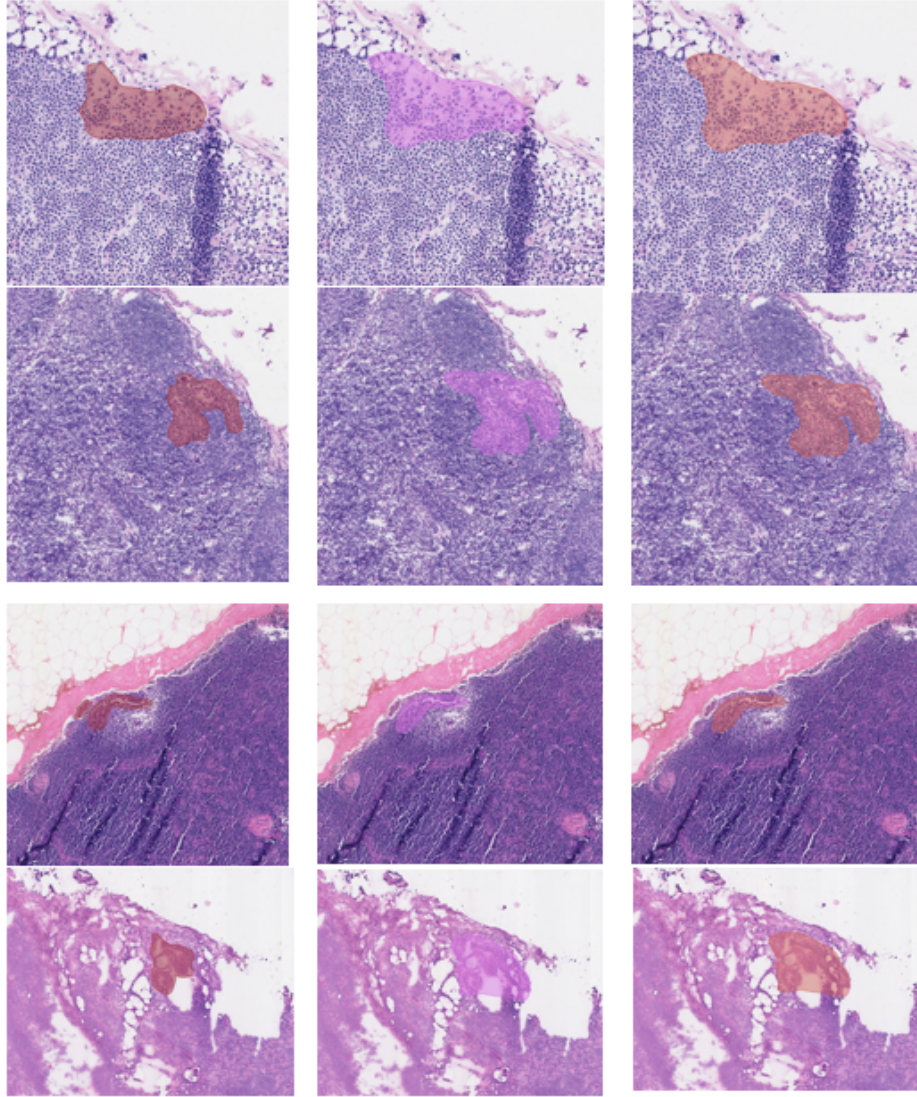


Figure 1: Qualitative comparison of segmentation results on the challenging FS-Mic (intraoperative frozen section microscope) dataset. Each row displays a different clinical case. From left to right, the columns show the segmentation masks generated by PathChat-SegR1, MedSAM, and nnU-Net, respectively. Our model, PathChat-SegR1, consistently provides more accurate and detailed delineations of the metastatic regions. It excels in capturing complex boundaries and subtle features, whereas MedSAM and nnU-Net tend to produce less precise masks, often struggling with the artifacts and staining heterogeneity inherent in frozen section images.

A.9 SEMI-AUTOMATED ANNOTATION PIPELINE

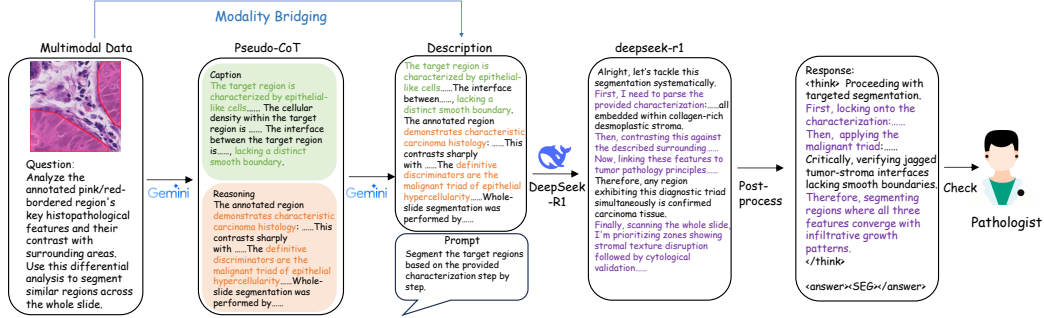


Figure 2: Illustration of our semi-automated pipeline for generating Chain-of-Thought (CoT) reasoning annotations. The process begins with a pseudo-CoT, which is refined by Gemini-2.5-Pro into a structured morphological description and a direct segmentation prompt. Subsequently, the DeepSeek model is employed to expand this structured information into a detailed, step-by-step reasoning path. Finally, all generated annotations undergo manual review and refinement by expert pathologists to ensure clinical accuracy and relevance.

A.10 PERFORMANCE RADAR CHART ACROSS MODELS

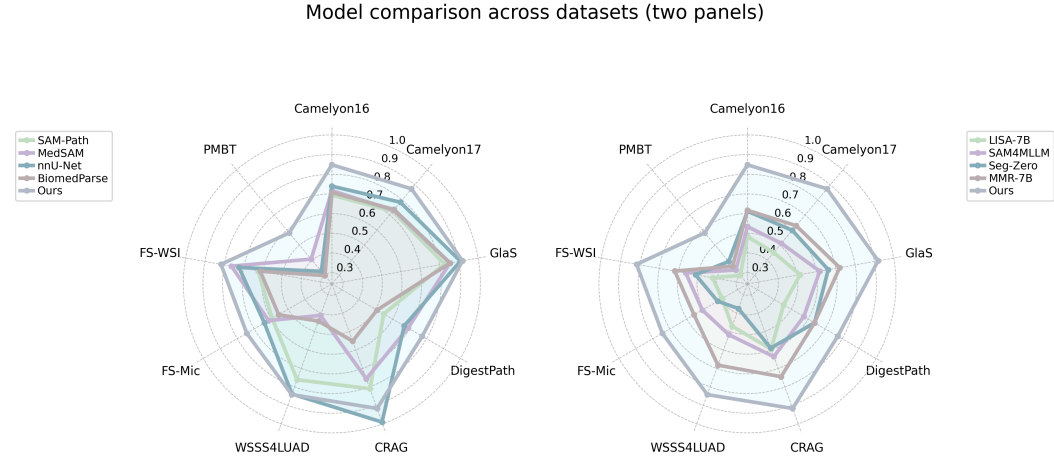


Figure 3: Comprehensive performance radar chart comparing PathChat-SegR1 against baseline models across the entire benchmark. The chart visualizes Dice scores on various datasets, where PathChat-SegR1 (green solid line) consistently demonstrates leading performance. Its exceptional generalization and robustness are particularly evident on challenging clinical datasets like intraoperative frozen sections (FS-Mic, FS-WSI) and zero-shot pulmonary metastasis from bone tumors (PMBT).

A.11 COMPREHENSIVE ABLATION STUDIES

Table 3: Performance ablation study on the FS-Mic dataset across different magnifications (40x, 20x, and 10x). The results indicate that PathChat-SegR1 has superior performance and robustness, particularly at lower magnifications where greater contextual understanding is required. In contrast, MedSAM’s performance degrades significantly at 10x, highlighting the limitations of models without explicit reasoning capabilities when faced with scale variations.

Magnification	MedSAM	PathChat-SegR1
40x	0.735	0.891
20x	0.691	0.812
10x	0.205	0.503
Avg	0.54	0.73

Table 4: Performance advantage of SO-GRPO on tasks of varying difficulty. The relative improvement is more pronounced on harder tasks.

Dataset / Scenario	Standard GRPO	SO-GRPO	Relative Improvement
Easy (Camelyon16)	0.74	0.76	+2.7%
Medium (CRAG)	0.88	0.92	+4.5%
Hard (FS-Mic)	0.68	0.74	+8.8%
Very Hard (PMBT zero-shot)	0.53	0.58	+9.4%
Extreme (Rare Disease)	0.45	0.53	+17.8%

Meaningful and Synergistic Improvements. As shown in Table ??, standard GRPO is already a strong baseline, improving the Dice score from 0.40 (SFT only) to 0.53. However, our proposed SO-GRPO pushes performance further to 0.58. This 5-point improvement is of significant clinical value, especially in the zero-shot context. The ablation study shows that each component provides an incremental contribution: GAE (+2 points), soft rewards (+2 points), and the sparsity constraint (+1 point) work in concert, producing a result that exceeds the simple sum of their parts and demonstrates a positive synergistic effect.

The Harder the Task, the Greater the Advantage. Crucially, the benefits of SO-GRPO are most pronounced when tackling complex tasks. As clearly illustrated in Table 4, the performance gain of SO-GRPO over standard GRPO widens as task difficulty increases. The improvement ranges from 2.7% on the relatively simple Camelyon16 dataset to a remarkable 17.8% on the extremely challenging rare disease zero-shot task. This strongly validates that the optimizations within SO-GRPO are essential for navigating the complex reasoning and uncertainty inherent in difficult, real-world clinical scenarios.

Improved Stability and Efficiency. The performance gains of SO-GRPO are not only rooted in final segmentation accuracy but also in the quality of its reasoning process and its training efficiency. As detailed in Table 5, SO-GRPO generates the `<SEG>` token with a 62.7% lower false positive rate and 24.0% better positional accuracy. Furthermore, Table 6 shows that the SO-GRPO training process is more stable (35.4% lower gradient variance) and converges 25% faster than standard GRPO, achieving a higher level of performance in fewer training steps.

Table 5: Analysis of `<SEG>` token generation quality. SO-GRPO exhibits higher precision and stability in the reasoning-decision phase.

Metric	Standard GRPO	SO-GRPO	Improvement
Precision	0.62	0.67	+8.1%
Recall	0.73	0.75	+2.7%
Position accuracy (pixels)	31.2	23.7	+24.0%
False positive rate	8.3%	3.1%	-62.7%
Training stability (std)	0.048	0.031	+35.4%

Table 6: Convergence efficiency analysis. SO-GRPO achieves higher performance in less time.

Metric	Standard GRPO	SO-GRPO	Analysis
Steps to 80% performance	12K	9K	25% faster
Steps to 90% performance	20K	15K	25% faster
Steps to convergence	24K	18K	25% faster
Final performance	0.53	0.58	9.4% higher
Performance@10K steps	0.46	0.51	Stronger early on

A.12 COMPLETE TRAINING AND REPRODUCIBILITY DETAILS

- **SO-GRPO Components** (addressing reviewer concerns):
 - Reward weights: $\lambda_1 = 0.3$ (Dice), $\lambda_2 = 0.2$ (IoU), selected via grid search over $[0.1, 0.5]$
 - GAE parameters: $\lambda = 0.95$, $\gamma = 0.99$ (standard RL values)
 - Sparsity control: $\beta = 0.01$, selected to achieve 15-20% SEG_i generation rate
 - KL penalty: $\lambda_{KL} = 0.01$ for stable training
 - Differentiable reward: $\epsilon = 10^{-7}$ for numerical stability
- **Model Configuration:**
 - LoRA rank: $r = 16$ (balancing performance and efficiency)
 - Learning rate: 1×10^{-4} (SFT), 5×10^{-5} (SO-GRPO)
 - Batch size: 64 (SFT), 32 (SO-GRPO)
 - Gradient accumulation: 2 steps
- **Hyperparameter Selection Process:**
 - Conducted grid search on validation set
 - Key finding: $\lambda_1 : \lambda_2$ ratio of 3:2 optimal for pathology images
 - Sparsity β critical: too high causes under-generation, too low causes over-generation