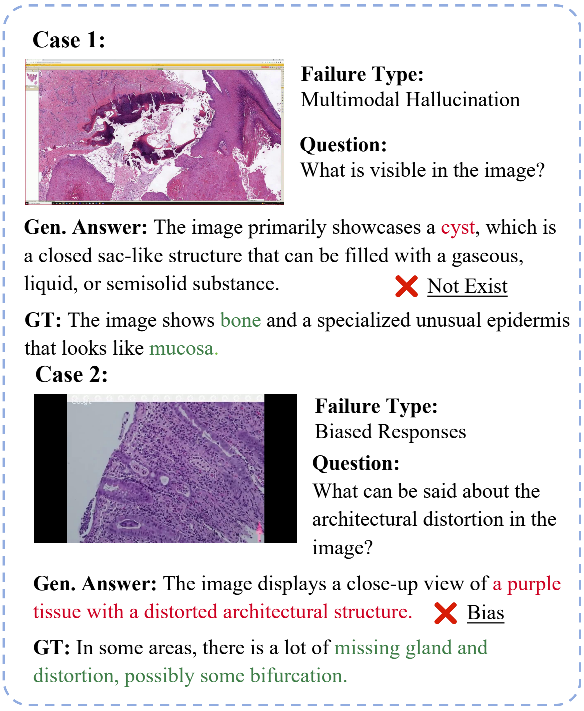
EAGLE: Expert-Guided Self-Enhancement for Preference Alignment in Pathology Large Vision-Language Model

Meidan Ding[1,2,3,†], **Jipeng Zhang**[4,†], **Wenxuan Wang**[5], **Haiqin Zhong**[6],
Xiaoqin Wang[1,2,3], **Xinheng Lyu**[1,7], **Wenting Chen**[8,*], **Linlin Shen**[1,2,3,*]
[1]Computer Vision Institute, College of Computer Science and Software Engineering, Shenzhen University
[2]School of Artificial Intelligence, Shenzhen University [8]City University of Hong Kong
[3]Guangdong Provincial Key Laboratory of Intelligent Information Processing
[4]The Hong Kong University of Science and Technology [5]Renmin University of China
[6]School of Biomedical Engineering, Shenzhen University [7]University of Nottingham Ningbo China

Abstract

Recent advancements in Large Vision Language Models (LVLMs) show promise for pathological diagnosis, yet their application in clinical settings faces critical challenges of multimodal hallucination and biased responses. While preference alignment methods have proven effective in general domains, acquiring high-quality preference data for pathology remains challenging due to limited expert resources and domain complexity. In this paper, we propose **EAGLE** (**E**xpert-guided self-enhancement for preference **A**lignment in patholo**G**y **L**arge vision-languag**E** model), a novel framework that systematically integrates medical expertise into preference alignment. EAGLE consists of three key stages: initialization through supervised fine-tuning, self-preference creation leveraging expert prompting and medical entity recognition, and iterative preference following-tuning. The self-preference creation stage uniquely combines expert-verified chosen sampling with expert-guided rejected sampling to generate high-quality preference data, while the iterative tuning process continuously refines both data quality and model performance. Extensive experiments demonstrate that EAGLE significantly outperforms existing pathological LVLMs, effectively reducing hallucination and bias while maintaining pathological accuracy. The source code is available at https://github.com/meidandz/EAGLE.



Case 1:

Failure Type: Multimodal Hallucination

Question: What is visible in the image?

Gen. Answer: The image primarily showcases a cyst, which is a closed sac-like structure that can be filled with a gaseous, liquid, or semisolid substance. ✗ Not Exist

GT: The image shows bone and a specialized unusual epidermis that looks like mucosa.

Case 2:

Failure Type: Biased Responses

Question: What can be said about the architectural distortion in the image?

Gen. Answer: The image displays a close-up view of a purple tissue with a distorted architectural structure. ✗ Bias

GT: In some areas, there is a lot of missing gland and distortion, possibly some bifurcation.

Figure 1: Typical failure cases demonstrate **multimodal hallucination** and **biased** response issues in the SFT-tuned model. Gen. Answer and GT represent the answer generated by the SFT model and ground-truth answer.

1 Introduction

Histopathological examination is widely regarded as the gold standard in cancer diagnosis (Zhu et al., 2024b), requiring meticulous analysis of tissue samples to detect abnormalities in cells and tissues (Gurcan et al., 2009). This process heavily relies on the expertise of pathologists to interpret subtle patterns indicative of various disease states (Van der Laak et al., 2021). However, these traditional diagnostic methods are often time-intensive (Pallua et al., 2020) and subject to subjective interpretation (Baxi et al., 2022), prompting the integration of artificial intelligence (AI) techniques to enhance diagnostic accuracy and efficiency.

Recent advancements in Large Vision Language Models (LVLMs) (Liu et al., 2024b; Zhu et al., 2023; Li et al., 2024a; Bai et al., 2023) have demonstrated remarkable performance across general domains. However, the application of these capabilities to medical area, particularly in pathology, demands great attention to reliability and safety

---

[†] These authors contributed equally.

[*] Corresponding Author: wentichen7-c@my.cityu.edu.hk, llshen@szu.edu.cn

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14603–14619
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

due to stringent ethical considerations (Liu et al., 2024c). While LVLMs like LLaVA show promising improvements through Supervised Fine-Tuning (SFT) with pathology-specific instructions (Seyfioglu et al., 2024; Sun et al., 2024; Dai et al., 2024), significant challenges persist in achieving robust alignment with human expertise. Specifically, there are two aspects as Fig. 1 shows: **1) Multi-modal Hallucination:** A key challenge in pathological AI is hallucination—descriptions or diagnoses that misalign with image content (Li et al., 2024c, 2023; Zhao et al., 2023; Xie et al., 2024b). Unlike general vision-language tasks, pathology hallucinations stem from subtle misidentifications of morphological patterns, potentially leading to serious diagnostic errors. For example, as shown in Fig. 1, a bone pathology image was mistakenly identified as a cyst, highlighting the risks of such misinterpretations in clinical practice. **2) Biased Responses:** Models tend to default to common diagnoses, overlooking rare pathological conditions due to dataset imbalances. This bias reduces accuracy in less frequent cases, posing a significant challenge in pathology where rare diseases require precise identification for proper treatment. In view of this, solving the above two problems is particularly important to further improve the robustness and stability of the LVLMs for pathology. While Seyfioglu et al. (2024) proposes Quilt-LLaVA, a specific pathological LVLM fine-tuned by a pathology instruction dataset, it cannot mitigate the alignment issues above. Based on the successful experience in the general domain (Rafailov et al., 2024; Li et al., 2024c; Xiong et al., 2024; Zhao et al., 2023), the preference alignment method (PFT) is a potential way to mitigate these issues.

When designing the alignment method, we identify significant challenges in acquiring high-quality multi-modal preference data for pathology. Traditional methods of preference data annotation, including human expert annotation (Zhou et al., 2024) and proprietary-level models annotation (Taori et al., 2023; Chiang et al., 2023a), face substantial limitations in this context. **Firstly**, scarce pathology data coupled with the high costs of expert annotation create fundamental barriers to acquiring the amount of data required for the model alignment. While proprietary-level models present a potentially more scalable alternative, they frequently lack the specialized domain knowledge essential for accurate pathological assessment. Meanwhile, proprietary-level models often require API calls and data uploads, which introduce the risk of data leakage. **Secondly**, current preference data generation methods lack integration of critical medical domain knowledge that is essential for accurate pathological diagnosis. Pathology is heavily dependent on specialized expertise, including understanding of tissue morphology patterns, disease progression mechanisms, rare pathological presentations, and complex clinical correlations. Without incorporating this domain-specific knowledge, generated preference data may not capture nuanced decision-making processes that pathologists use, potentially leading to suboptimal model performance. Furthermore, due to the complexity of pathological knowledge and the open-ended nature of diagnostic reasoning, it becomes difficult to verify the accuracy of both chosen and rejected responses in the preference dataset. This verification challenge raises concerns about the quality and reliability of the training data. It is necessary to develop more robust approaches to preference data annotation that effectively incorporate medical expertise while ensuring verifiable quality in pathological LVLMs.

Table 1: Comparison among SFT, PFT, and EAGLE.

	Preference Alignment	Data Self-Creation	Iterative Improvement	Expertise
SFT	×	×	×	×
PFT	✓	×	×	×
EAGLE	✓	✓	✓	✓

To address these challenges, we propose EAGLE (**E**xpert-guided self-enhancement for preference **A**lignment in patholo**G**y **L**arge vision-languag**E** model), a novel framework to effectively integrates medical domain knowledge into preference alignment through automated preference data generation, comprising three key stages: initialization, self-preference creation, and preference following-tuning stages. In Table 1, EAGLE uniquely combines preference alignment, self-created training data, iterative improvement, and domain expertise integration. In the *initialization* stage, we establish the model's basic pathological understanding through supervised fine-tuning on instruction-following data. The core innovation lies in our *self-preference creation* stage, where we integrate medical domain knowledge into preference data generation through: (1) an expert-verified chosen sampling module that leverages expert prompting for high-quality responses, and

(2) an expert-guided rejected sampling module that incorporates multiple pathology expert models and medical entity recognition to ensure rejected responses maintain pathological coherence while containing controlled inaccuracies. After this stage, we construct a preference dataset with about 150k paired data for preference alignment. Finally, our *preference following-tuning* stage iteratively aligns the model with these expert-informed preference pairs through direct preference optimization. This iterative process involves continuous updates to both the preference data and the model parameters, gradually enhancing both the quality of the preference data and the overall model performance. Extensive experiments demonstrate the superiority of EAGLE to existing pathological LVLMs. Our contributions can be summarized as follows:

- We propose EAGLE, a novel three-stage framework that addresses the challenges of pathology LVLM alignment through expert-guided preference optimization. To our knowledge, this is the first work to systematically integrate medical expertise into preference alignment for pathological LVLMs.
- The proposed preference dataset contains more than 150k preference data, which can be used for preference alignment, helping the model to solve problems such as multimodal hallucination and biased responses.
- Experimental results on different models show the superiority of the proposed EAGLE over existing pathological LVLMs.

2 Related Work

2.1 Medical Large Vision Language Models (LVLMs)

Recent LLM advancements have transformed vision-language models (Li et al., 2024a; Yin et al., 2024; Gu et al., 2024), with models like LLaVA (Liu et al., 2024b) achieving GPT-4-level performance. In medicine, LLaVA-Med (Li et al., 2024b) and BiomedGPT (Zhang et al., 2024) pioneered innovations, with BiomedGPT training on 271,803 image-text pairs. While MedDr (He et al., 2024) and STLLaVA-Med (Sun et al., 2024) advanced healthcare applications, histopathology remained underserved. PathChat (Lu et al., 2023), adapted from LLaVA and fine-tuned on 450k instruction pairs, later outperformed GPT-4V in pathology tasks. Quilt-LLaVA (Seyfioglu et al., 2024), trained on the Quilt-instruct dataset and Quilt-VQA's 107,131 QA pairs, further advanced pathology capabilities. Meanwhile, Xie et al. proposed the MedTrinity-25M dataset and used this dataset to pretrain LLaVA, surpassing representative SOTA multimodal large language models. However, challenges like high construction costs and alignment issues persist. Thus, we propose a self-creation stage to synthesize preference data automatically.

2.2 Preference Learning in LVLM

Preference learning is essential for developing effective (Gorbatovski et al., 2024), safe (Dai et al., 2023), and trustworthy (Yu et al., 2024) models. It typically follows an instruction-following stage, where models are fine-tuned using human input. Cui et al. propose a data-centric framework, Biomedical Visual Instruction Tuning with Clinician Preference Alignment (BioMed-VITAL), that incorporates clinician preferences into both stages of generating and selecting instruction data for tuning biomedical multimodal foundation models. Reinforcement Learning with Human Feedback (RLHF) (Ouyang et al., 2022) is a widely used method, involving a reward model trained on human preferences to guide policy optimization. Since RLHF is complex and unstable, requiring reward model training and fine-tuning of large unsupervised models to maximize rewards, several RLHF variants have been developed, including Direct Policy Optimization (DPO) (Rafailov et al., 2024), Proximal Policy Optimization (PPO) (Schulman et al., 2017), and Reinforcement Learning from AI Feedback (RLAIF) (Bai et al., 2022). DPO simplifies the process by bypassing the reward model and directly learning policies, reducing complexity and often outperforming PPO. Zhu et al. propose MMedPO, a novel multimodal medical preference optimization approach that considers the clinical relevance of preference samples to enhance Med-LVLM alignment. In this work, we propose EAGLE to iteratively align specially for pathology LVLMs with preference data optimized by DPO.

3 Methods

In Fig. 2, we propose EAGLE (**E**xpert-guided self-enhancement for preference **A**lignment in patholo**G**y **L**arge vision-languag**E** model), a novel

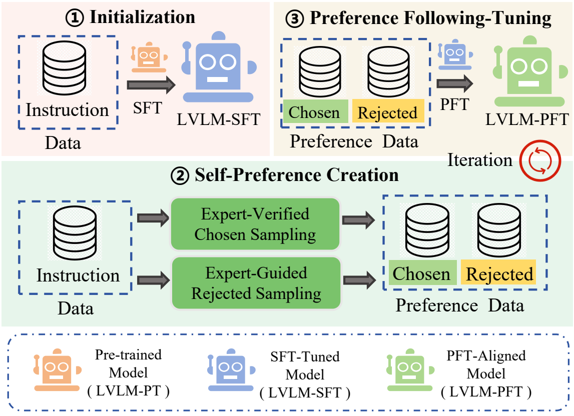

① Initialization
Instruction Data
SFT
LVLM-SFT
③ Preference Following-Tuning
Chosen
Rejected
Preference Data
PFT
LVLM-PFT
Iteration
② Self-Preference Creation
Instruction Data
Expert-Verified Chosen Sampling
Expert-Guided Rejected Sampling
Chosen
Rejected
Preference Data
Pre-trained Model (LVLM-PT)
SFT-Tuned Model (LVLM-SFT)
PFT-Aligned Model (LVLM-PFT)

Figure 2: Overview of EAGLE, a pathology LVLM framework integrating expert knowledge for preference alignment through three stages, i.e. (1) initialization for initial instruction tuning, (2) self-preference creation for expert-guided preference data generation, and (3) preference following-tuning for iterative alignment, with stages 2 and 3 implemented cyclically.

framework that leverages expert knowledge for efficient preference data generation in pathological analysis. It consists of three key stages, i.e., *Initialization*, *Self-Preference Creation*, and *Preference Following-Tuning* stages. Firstly, the initialization stage trains a seed LVLM through SFT on instruction-following datasets to obtain an LVLM-SFT model with basic pathological understanding and interaction capabilities. Then, a self-preference creation stage utilizes an expert-verified chosen sampling and expert-guided rejected sampling modules to generate preference data, i.e. chosen and rejected samples, respectively. With the chosen and rejected samples, the preference following-tuning stage leverages DPO to align the LVLM-SFT model with the constructed preference data to obtain the LVLM-PFT model. The second and third stages are implemented iteratively to improve the preference alignment, as shown in Algorithm 1.

3.1 Initialization

The initialization stage creates a basic LVLM through supervised fine-tuning on pathological instruction data from (Seyfioglu et al., 2024). The model combines a Vicuna-based LLM (Chiang et al., 2023b) ($f_\phi(\cdot)$) with a QuiltNet (Ikezogwo et al., 2024) visual encoder, connected via a projection layer. Joint optimization of these components using the instruction dataset produces an initial LVLM-SFT model π_θ capable of basic pathological query interpretation.

Algorithm 1 EAGLE Pipeline

```
Input: Instruction following data
Output: Enhanced LVLM $\pi_{\text{pre}}$
Stage 1: Initilization
IFT_data ← Construct_Instruction_Data()
$\pi_\theta$ ← Supervised_Fine_Tuning(IFT_data)
Stage 2: Self-Preference Creation
Preference_data ← Create_Preference_Dataset($\pi_\theta$)
Stage 3: Preference Following-Tuning (PFT)
$\pi_{\text{pre}}$ ← PFT(Preference_data)
while Update and Refine do
    $\pi_\theta \leftarrow \pi_{\text{pre}}$
    Stage 2
    Stage 3
end while
```

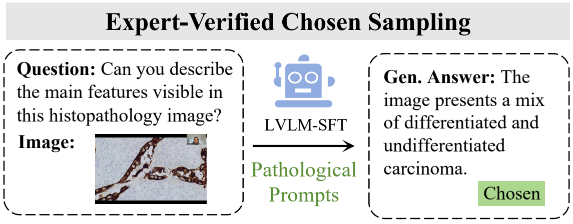

Expert-Verified Chosen Sampling
Question: Can you describe the main features visible in this histopathology image?
Image:
LVLM-SFT
Pathological Prompts
Gen. Answer: The image presents a mix of differentiated and undifferentiated carcinoma.
Chosen

Figure 3: Expert-verified chosen sampling to generate chosen responses through pathological prompts.

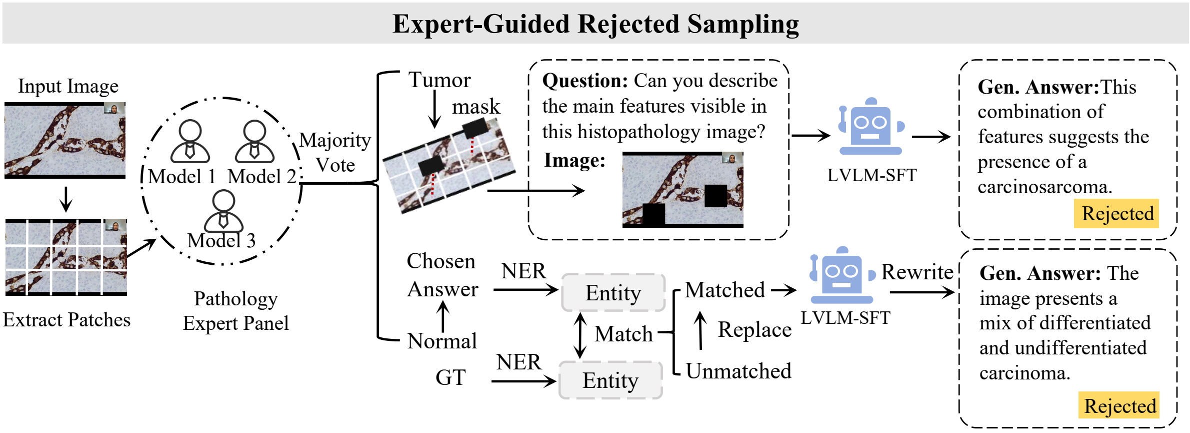
3.2 Self-Preference Creation

Due to the scarcity of preference data in pathology, we introduce a novel Self-Preference Creation stage to leverage expert knowledge to efficiently generate high-quality preference data at a low cost, avoiding time-consuming manual annotations from medical professionals. It consists of an expert-verified chosen sampling and expert-guided rejected sampling modules, which generate the chosen and rejected responses respectively to construct the preference training data. Formally, the preference dataset $\mathcal{D}$ consists of N samples, where each sample contains an image $\mathbf{X}_i$ and a question $\mathbf{Q}_i$ paired with two responses: a chosen response $\mathbf{A}_i^c$ and a rejected response $\mathbf{A}_i^r$:

$$\mathcal{D} = \{\mathbf{Q}_i, \mathbf{A}_i^c, \mathbf{A}_i^r\}_{i=1}^N, \tag{1}$$

where $\mathbf{A}_i^c$ is generated through expert-verified chosen sampling and $\mathbf{A}_i^r$ through expert-guided rejected sampling.

Expert-Verified Chosen Sampling (EVCS). In Fig. 3, the expert-verified chosen sampling module is designed to generate high-quality responses for pathological diagnosis. We use the LVLM-SFT π_θ



Question: Can you describe the main features visible in this histopathology image?

Gen. Answer: This combination of features suggests the presence of a carcinosarcoma.

Gen. Answer: The image presents a mix of differentiated and undifferentiated carcinoma.

Figure 4: Self-preference creation stage to generate rejected responses with expert guidance.

to generate the chosen response $\mathbf{A}_i^c$ by prompting the model to "role-play" as a domain expert,

$$\mathbf{A}_i^c = \pi_\theta(\mathbf{X}_i, \mathbf{Q}, \mathbf{P_c}), \tag{2}$$

where $\mathbf{P_c}$ is a crafted prompt template that instructs the model to adopt the persona of a knowledgeable expert in pathological diagnosis. Please see Appendix D for details.

Expert-Guided Rejected Sampling (EGRS). In Fig. 4, we propose an expert-guided rejected sampling module to utilize multiple experts to analyze the pathological findings of the input image and further generate rejected responses based on the findings.

Concretely, given an input image $\mathbf{X}_i$, we extract patches for $\mathbf{X}_i$ and establish a pathology expert panel to screen the abnormality of each patch $\mathbf{X}_i^j$, where j indicates the j-th patch of the input image. The pathology expert panel consists of four pathological expert models, including UNI (Chen et al., 2024), CTransPath (Wang et al., 2022), Enhanced ViT (Ding et al., 2023), and Phikon (Filiot et al., 2023). These models are either trained on large-scale pathology datasets or trained for tumor/no-tumor binary classification, showing strong representation ability and zero-shot generalization capability. Each expert model provides the decision on the presence of tumors for each patch. To provide overall pathological findings, we perform a majority vote on the decisions from all the expert models. Then, we generate the rejected response based on the presence of the tumor. When the input image contains tumor regions, we generate a mask image $\mathcal{M}$ that indicates the tumor patches, and perform element-wise multiplication between $\mathcal{M}$ and $\mathbf{X}_i$ to generate a masked input image ($\mathcal{M} \odot \mathbf{X}_i$). With the masked image, LVLM-SFT generates a rejected response:

$$\mathbf{A}_i^r = \pi_\theta(\mathcal{M} \odot \mathbf{X}_i, \mathbf{Q}, \mathbf{P_r}), \tag{3}$$

where prompt $\mathbf{P_r}$ is designed to guide the model in generating responses that maintain pathological relevance while introducing controlled semantic deviations. Please see Appendix D for details.

For images without tumor regions, we first apply a Medical Named Entity Recognition (NER) model (Mattupalli, 2025) to identify pathological entities in both the chosen response and the ground-truth (GT) answer. We then perform exact matching between the entities of the chosen response $\mathbf{E}^c$ and GT answer $\mathbf{E}^{gt}$ to identify matched and unmatched entities: $\mathbf{E}^c = \{\mathbf{E}_m^c, \mathbf{E}_u^c\}$, $\mathbf{E}^{gt} = \left\{\mathbf{E}_m^{gt}, \mathbf{E}_u^{gt}\right\}$, where m and u denote 'matched' and 'unmatched', respectively. To generate the rejected response, we replace matched entities $\mathbf{E}_m^c$ from the chosen response with unmatched entities $\mathbf{E}_u^{gt}$ from the GT answer, creating a synthetic response that introduces controlled inaccuracies while maintaining pathological coherence. These perturbations are carefully designed to reflect realistic diagnostic errors, ensuring that the rejected samples maintain a high degree of relevance to real-world scenarios. Finally, we refine this synthetic response through LVLM-SFT to improve its coherence, yielding the final rejected response. This approach ensures rejected responses contain plausible but incorrect diagnostic information, providing effective training signals for preference optimization.

By adopting the self-preference creation stage, we construct a preference dataset with more than 150k paired data for preference alignment, helping the model to solve problems such as multimodal hallucination and biased responses.

3.3 Preference Following-Tuning

Aiming to align the SFT-tuned LVLM (LVLM-SFT) with the synthetic preference data, we introduce Preference Following-Tuning (PFT) to learn from the preference data iteratively. Specifically, we adopt Direct Preference Optimization (DPO) (Rafailov et al., 2024) to implement preference alignment, which can be viewed as a new parameterization of the reward model in RLHF (Reinforcement Learning from Human Feedback) (Ouyang et al., 2022) that can directly optimize the policy model π_{pre}. The policy model is optimized as follows:

$$\begin{aligned} L_{dpo}(\pi_{pre};\pi_\theta) = -&E_{(m,\mathbf{A}_i^c,\mathbf{A}_i^r)\sim\mathcal{D}} \\ &\Big[\log\sigma(\beta log\frac{\pi_{pre}(\mathbf{A}_i^c|m)}{\pi_\theta(\mathbf{A}_i^c|m)} \\ &-\beta\log\frac{\pi_{pre}(\mathbf{A}_i^r|m)}{\pi_\theta(\mathbf{A}_i^r|m)})\Big], \end{aligned} \tag{4}$$

where $m = (\mathbf{X}_i, \mathbf{Q})$, and is the input to the LVLM π_{pre}. π_{pre} and π_θ denote policy model and reference model, respectively. $\mathcal{D}$ represents our preference dataset, and logσ stands for log-sigmoid function. This objective function is designed to train the reward and policy models concurrently, skewing the reward model to favor the chosen response and reject the rejected response.

To enhance the quality of preference alignment, we implement stages 2 and 3 alternately, allowing iterative refinement of both chosen and rejected responses using the progressively improved LVLM-PFT model. This creates a positive feedback loop where preference dataset evolves alongside model improvements, leading to enhanced response quality and more effective preference alignment.

4 Experiments

4.1 Implementation Details

For Supervised Fine-Tuning (SFT), we adopt deepspeed ZeRO stage-3 (Liu et al., 2024a) during SFT. The total epoch, batch size, learning rate, and weight decay are set as 1, 128, 2e-4, and 0, respectively. The hyper-parameters of lora γ and α are set as 128, 256. For Preference Following-Tuning (PFT), we adopt deepspeed ZeRO stage-3 (Liu et al., 2024a) during PFT with DPO. The total epoch, batch size, learning rate, and weight decay are set as 1, 128, 2e-6, and 0, respectively. The hyper-parameters of lora γ and α are set as 128, 256, and the scale parameter β in DPO is fixed as 0.1. During these stages, the visual encoder is frozen, and the project layer and LLM are trainable.

4.2 Datasets

We evaluate EAGLE on pathological datasets: PathVQA(He et al., 2020), Quilt-VQA (Q-VQA), Quilt-Red (Q-Red), Quilt-Nored (Q-NoRed), and Quilt-conversation (Seyfioglu et al., 2024), across three tasks: Visual Question Answering (VQA), VQA with visual prompting, and multi-modal conversation. PathVQA and Q-VQA assess VQA performance, Q-Red and Q-NoRed evaluate visual prompting, and Quilt-conversation tests multi-modal conversations.

PathVQA (He et al., 2020) contains 32,799 question-answer pairs extracted from 4,998 pathology image-caption pairs sourced from textbooks and digital libraries. For our evaluation, we utilized 3,370 open-instructions from the evaluation set.

Quilt-VQA (Seyfioglu et al., 2024) consists of 1,283 pairs which are categorized into two groups, 1,055 image-dependent pairs that reference the narrator's specific image and 228 general-knowledge pairs rooted in broader medical understanding.

Quilt-Red (Q-Red) and Quilt-Nored (Q-NoRed) (Seyfioglu et al., 2024) are construct based on Quilt-VQA. Quilt-Red (Q-Red) is constructed by selecting a subset of Quilt-VQA, and creating red ellipses in the input image to provide visual guidance for the question. The instructions in Q-Red contain 'Tip: Focus on image areas highlighted by the red circle(s)', while Quilt-Nored (Q-NoRed) does not contain this instruction and red ellipses in the input image. They both have 426 images and 439 instructions.

Quilt-conversation (Seyfioglu et al., 2024) comprises 14 subtypes and 326 questions, including 256 conversations and 70 detailed descriptions, as detailed in Table 10. It is named instruction following test set in the Quiltllava work (Seyfioglu et al., 2024).

4.3 Results on Visual Question Answering

Quantitative results. Table 2 presents a comprehensive comparison between our EAGLE and existing LVLMs, including LLaVA (Liu et al., 2024b) for general scenes, LLaVA-Med (Li et al., 2024b), STLLaVA-Med(Sun et al., 2024), MMedPO (Zhu

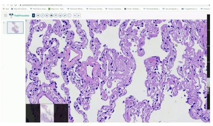
Table 2: Quantitative evaluation of VQA and VQA with visual prompting (VQA w/ V.P.) tasks. Q-VQA, Q-Red, Q-NoRed, and Avg denote Quilt-VQA, Quilt-Red, Quilt-NoRed, and average recall, respectively.

LVLMs	VQA (%)			VQA w/ V.P.(%)			Avg
	PathVQA	Q-VQA	Avg	Q-Red	Q-NoRed	Avg	
LLaVA	11.68	53.66	32.67	55.48	55.01	55.25	43.96
LLaVA-Med	8.20	49.63	28.92	50.11	47.91	49.01	38.97
STLLaVA-Med	12.06	56.60	34.33	58.23	54.72	56.48	45.41
Quilt-LLaVA	13.24	54.62	33.93	55.68	52.84	54.26	44.10
MMedPO	8.94	48.94	28.94	42.35	47.79	45.07	37.00
MMed-RAG	8.03	48.71	28.37	50.16	48.33	49.24	38.80
EAGLE	**15.80**	**66.42**	**41.11**	**64.95**	**64.21**	**64.58**	**52.85**

et al., 2024a) and MMed-RAG (Xia et al., 2024) for medical scenes, and Quilt-LLaVA (Seyfioglu et al., 2024) for pathology-specific scenes. Our EAGLE achieves superior performance with an average recall of 41.11% on PathVQA and Quilt-VQA (Q-VQA) datasets, surpassing the second-best performer STLLaVA-Med by 6.78%. Despite PathVQA being a challenging dataset with low overall performance across all models, EAGLE still achieves the best result of 15.80%. On Quilt-VQA, it also demonstrates strong capability with 66.42% recall, validating its effectiveness in understanding and answering questions about pathological images.

Qualitative results. In Table 3, we present a comparative analysis of different LVLMs on a histopathology VQA example. When asked about NSIP characteristics, LLaVA completely misinterprets the condition, incorrectly associating it with liver disease rather than lung pathology. Quilt-LLaVA shows improvement by correctly identifying NSIP as a lung condition and recognizing its uniform appearance, though it still includes some imprecise descriptions of pathological features. In contrast, EAGLE demonstrates superior understanding by providing a comprehensive and accurate description of NSIP's histopathological features, particularly emphasizing the uniform pattern of interstitial changes and correctly differentiating it from UIP, which aligns closely with the ground truth. This shows EAGLE's enhanced capability in interpreting complex medical images and providing more precise, clinically relevant responses compared to existing approaches. Please see Appendix Table 9 for more details.

Table 3: Qualitative comparison of different LVLMs in VQA task on QuiltVQA dataset. Green for correct and red for wrong.

Input example, Histopathology Image and Question:	
User	 What characterizes the cases of NSIP?
LLaVA	The cases of NSIP (Nonalcoholic Steatohepatitis) are characterized by the presence of inflammation and fibrosis in the liver.
Quilt-LLaVA	In the cases of Non-Specific Interstitial Pneumonia (NSIP), there are several key features to note. NSIP is characterized by a uniform or 'non-patchy' appearance, which is a distinguishing factor from other interstitial lung diseases. This uniformity is seen in the image as a consistent pattern of interstitial inflammation and fibrosis.
EAGLE	The histopathological features of NSIP can vary widely, but they typically include a uniform pattern of interstitial inflammation and fibrosis, with no significant architectural distortion or honeycombing. The fibrosis in NSIP is usually uniform and lacks the characteristic features of usual interstitial pneumonia (UIP), such as a patchy distribution and fibroblastic foci.
Ground-Truth	NSIP cases are characterized by uniform expansion of the alveolar septa. It's a very homogeneous process in contrast to cases of UIP.

4.4 Results on VQA with Visual Prompting

We evaluate our method in VQA with the visual prompting task on the Quilt-Red (Q-Red) and Quilt-Nored (Q-NoRed) (Seyfioglu et al., 2024) datasets. The Q-Red dataset contains images with red elliptical annotations that highlight key contextual regions. In contrast, Q-NoRed images lack these visual guides, presenting a more challenging scenario for visual prompting.

Quantitative results. In Table 2, we quantitatively evaluate our EAGLE and other LVLMs on VQA with visual prompting (VQA w/ V.P.) task. EAGLE demonstrates consistent superior performance across both Q-Red and Q-NoRed datasets, achieving 64.95% and 64.21% recall respectively. Most notably, the minimal performance gap (0.74%) between Q-Red and Q-NoRed scenarios indicates EAGLE's robust ability to understand and reason about visual content, even without explicit visual prompts. This is in stark contrast to previous methods, which typically show larger performance degradation when visual guides are removed. For instance, STLLaVA-Med drops by 3.51% and

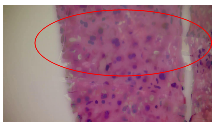
Quilt-LLaVA by 2.84%, suggesting that our approach particularly excels at bridging the gap between medical visual understanding and language reasoning. These results collectively demonstrate EAGLE's effectiveness in maintaining consistent performance across varying levels of visual guidance, indicating a more fundamental improvement in visual-language understanding rather than mere reliance on visual prompts.

Qualitative results. Table 4 compares EAGLE with existing LVLMs on the Q-Red dataset. LLaVA's response shows limited understanding, incorrectly identifying bile ducts where none are prominently featured. While Quilt-LLaVA provides a more detailed explanation and correctly identifies bile accumulation, it includes inaccurate observations about hepatocyte appearance. EAGLE not only correctly identifies the cholestatic pattern but also accurately associates it with drug toxins, aligning closely with the ground truth. This exemplifies EAGLE 's exceptional vision-language understanding capabilities, where the iterative preference training enables the model to capture subtle pathological features and provide nuanced medical interpretations that are both accurate and clinically relevant. Please see Appendix Table 11 for details.

4.5 Results on Multi-Modal Conversation

We evaluate EAGLE and other LVLMs in multi-modal conversation task on the Quilt-conversation dataset, and use relative score (Seyfioglu et al., 2024) and *Elo* rating (Chiang et al., 2024) as metrics. In the relative score, Table 5 shows EAGLE achieves the best performance with an overall score of 69.5, surpassing Quilt-LLaVA (68.8) and LLaVA (51.5). For conversational questions, EAGLE scores 72.7, improving over Quilt-LLaVA (72.0) and LLaVA (55.0). On descriptive questions, it maintains competitive performance at 58.8, comparable to Quilt-LLaVA (58.4) and significantly above baselines. Across medical sub-domains, EAGLE demonstrates consistent superiority, with highest scores in Bone (81.0), Gynecologic (76.1), and Head and Neck (75.6) pathology. Notable improvements appear in challenging areas like Neuropathology (67.9 vs Quilt-LLaVA's 61.7). Detailed results are in Appendix Table 10. These results validate EAGLE's effectiveness in both general multi-modal conversation and specialized medical understanding. In *Elo* rating, we compare responses generated by different LVLMs and record the preferred reponses via GPT-4, which gives us a record of the winning rate between pairs of different LVLMs. In Fig. 5, EAGLE wins in the pairwise comparison with a score of 1,079, suggesting its superority to existing methods.

Table 4: Qualitative comparison in VQA with the visual prompting task on Q-Red dataset. Green for correct and red for wrong.

Input example, Histopathology Image and Question:	
User	 What is the pattern of cholestatic hepatitis associated with ? Tip: Focus on image areas highlighted by the red circle(s).
LLaVA	The pattern of cholestatic hepatitis associated with the red circle(s) in the image is characterized by the presence of numerous small bumps or bile ducts.
Quilt-LLaVA	The highlighted areas in the image, as indicated by the red circle(s), show a pattern of cholestatic hepatitis. This condition is characterized by the accumulation of bile within the liver due to impaired bile flow. The hepatocytes in these areas may appear swollen and have a pale, granular cytoplasm. The presence of cholestasis can be inferred from the presence of these features.
EAGLE	The image shows a pattern of cholestatic hepatitis, which is characterized by the presence of bile within the hepatocytes. This may be related to drugs and drug toxins. The pattern of cholestatic hepatitis is typically associated with a variety of liver diseases.
Ground-Truth	The pattern of cholestatic hepatitis seen in the image is associated with drug and medication toxins, as well as hepatitis A and E, as these can produce cholestatic injury features.

4.6 Ablation Studies

To prove the effectiveness of each stage, we ablate stages 2 (self-preference creation) and 3 (preference following-tuning) of our EAGLE. Table 6 shows that removing stages 2 and 3 decreases the performance significantly across all the datasets, indicating their effectiveness. To further analyze the effectiveness of the number of iterations (#Iteration) for stages 2 and 3, we compare the performance of EAGLE from 1 to 4 iterations. As shown in Table 6, performance improves in the first 2 iterations, notably on PathVQA and Q-VQA. Marginal gains are seen in the 3rd iteration for some tasks, while the 4th leads to overall declines. This suggests that stages 2 and 3 can gradually improve the LVLMs in pathological tasks through

Table 5: Quantitative comparison on the Quilt-conversation dataset, using the relative score via GPT-4 evaluation. Question type includes conversation (Conv) and detailed description (Desc). Sub-domain contains 14 domains. The number of each sub-domain is shown in brackets.

LVLMs	**Question Types**		**Sub-Domains**									**Overall**
	Conv	Desc	Bone	Breast	Cyto	Gastro	Gyne	H&N	Neuro	Pulm	Soft	
	(256)	(70)	(25)	(23)	(23)	(23)	(22)	(24)	(24)	(25)	(25)	
LLaVA (Liu et al., 2024b)	55.0	40.0	51.5	55.2	44.8	47.9	58.3	56.4	59.4	45.1	54.2	51.5
LLaVA-Med (Li et al., 2024b)	61.8	42.9	58.5	63.5	51.7	50.3	57.8	64.1	58.2	53.2	61.0	57.4
STLLaVA-Med (Sun et al., 2024)	69.1	47.9	69.4	67.1	50.9	63.9	65.8	73.2	63.5	53.7	69.4	64.2
Quilt-LLaVA (Seyfioglu et al., 2024)	72.0	58.4	77.4	75.1	59.8	59.3	61.0	70.4	61.7	56.4	80.5	68.8
EAGLE	**72.7**	**58.8**	**81.0**	**75.8**	**60.3**	**59.7**	**76.1**	**75.6**	**67.9**	**59.6**	**81.0**	**69.5**

Abbreviation: Cytopathology and Hepatopathology (Cyto), Gastrointestinal (Gastro), Gynecologic (Gyne), Head and Neck (H& N), Neuropathology (Neuro), Pulmonary (Pulm), and Soft tissue (Soft).

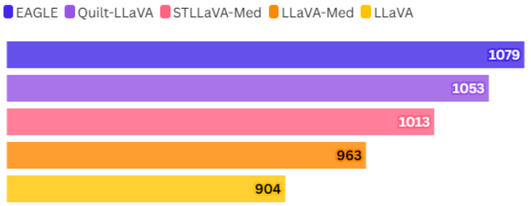


Figure 5: *Elo* rating comparison among LVLMs.

Table 6: Ablation Studies. Q-VQA, Q-Red, and Q-NoRed represent Quilt-VQA, Quilt-Red, and Quilt-NoRed, respectively.

Stage 1	Stage 2&3	#Iteration	Datasets				Average
			PathVQA	Q-VQA	Q-Red	Q-NoRed	
QuiltNet-B-32, Vicuna 7B Language Model							
✓			13.53	55.84	49.37	52.78	42.88
✓	✓	1	13.96	64.14	63.43	62.07	50.90
✓	✓	2	**15.80**	66.42	64.95	64.21	52.85
✓	✓	3	**15.80**	**66.89**	**65.33**	**64.53**	**53.13**
✓	✓	4	13.24	65.22	62.85	62.60	50.97
QuiltNet-B-32, Mistral 7B Language Model							
✓			12.61	55.40	53.24	43.99	41.31
✓	✓	1	13.42	56.50	61.45	60.24	47.90
✓	✓	2	13.96	57.02	62.27	63.24	49.14

Table 7: Performance under different sampling strategies.

EVCS	EGRS	Datasets				Average
		PathVQA	Q-VQA	Q-Red	Q-NoRed	
X	X	12.85	56.85	57.08	57.65	46.10
✓	X	13.05	62.54	62.85	61.26	49.92
✓	✓	**13.96**	**64.14**	**63.43**	**62.07**	**50.90**

iterations of preference training. The number of iterations is best controlled at 2-3 times, and too many iterations may cause overfitting. Moreover, we further compare our performance when using different language models, e.g. Vicuna 7B (Chiang et al., 2023b) and Mistral 7B (Jiang et al., 2023). The results show that the Vicuna 7B outperforms the Mistral 7B overall.

Table 7 shows the ablation study in different sampling strategies. When ablating EGRS by introducing random noise to generate rejected samples, the performance across all four datasets decreases significantly, highlighting the critical role of EGRS in generating high-quality rejected responses. While GT answers are available, we utilize LVLM-SFT to generate the "chosen" responses in preference pairs because the model, after supervised fine-tuning, is capable of producing clear and accurate responses that align closely with GT answers. Additionally, the model-generated answers often provide more context or explanatory detail than the GT, enhancing the quality of the "chosen" samples. The results also demonstrate the significant contributions of EVCS in preference data creation.

5 Conclusion

In this paper, we presented EAGLE, a novel framework for enhancing pathology-focused LVLMs through expert-guided preference alignment. EAGLE addresses multimodal hallucination and biased responses through a three-stage approach that generates high-quality preference data without extensive manual expert annotation. Extensive experiments across multiple pathology datasets demonstrate EAGLE's superiority over existing LVLMs, achieving the best performance in visual question answering, visual prompting tasks, and multimodal conversations.

Limitations

Although our method achieved promising experimental results, it was primarily designed for pathology applications. While the model demonstrates strong performance in pathology analysis, its generalization capabilities across different imaging modalities require enhancement. Future research should focus on expanding the model's adaptability to various medical imaging modalities and validating its cross-modal effectiveness.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant 82261138629 and 12326610; Guangdong-Macao Science and Technology Innovation Joint Fundation under Grant 2024A0505090003; Natural Science Foundation of Guangdong Province under Grant 2025A1515010184; the project of Shenzhen Science and Technology Innovation Committee under Grant JCYJ20240813141424032; and the Foundation for Young innovative talents in ordinary universities of Guangdong under Grant 2024KQNCX042.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Vipul Baxi, Robin Edwards, Michael Montalto, and Saurabh Saha. 2022. Digital pathology and artificial intelligence in translational medicine and clinical practice. *Modern Pathology*, 35(1):23–32.

Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. 2024. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023a. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. Blog post.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023b. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *Preprint*, arXiv:2403.04132.

Hejie Cui, Lingjun Mao, Xin Liang, Jieyu Zhang, Hui Ren, Quanzheng Li, Xiang Li, and Carl Yang. 2024. Biomedical visual instruction tuning with clinician preference alignment. *arXiv preprint arXiv:2406.13173*.

Dawei Dai, Yuanhui Zhang, Long Xu, Qianlan Yang, Xiaojing Shen, Shuyin Xia, and Guoyin Wang. 2024. Pa-llava: A large language-vision assistant for human pathology image understanding. *arXiv preprint arXiv:2408.09530*.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2023. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*.

Meidan Ding, Aiping Qu, Haiqin Zhong, Zhihui Lai, Shuomin Xiao, and Penghui He. 2023. An enhanced vision transformer with wavelet position embedding for histopathological image classification. *Pattern Recognition*, 140:109532.

Alexandre Filiot, Ridouane Ghermi, Antoine Olivier, Paul Jacob, Lucas Fidon, Axel Camara, Alice Mac Kain, Charlie Saillard, and Jean-Baptiste Schiratti. 2023. Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv*, pages 2023–07.

Alexey Gorbatovski, Boris Shaposhnikov, Alexey Malakhov, Nikita Surnachev, Yaroslav Aksenov, Ian Maksimov, Nikita Balagansky, and Daniil Gavrilov. 2024. Learn your reference model for real good alignment. *arXiv preprint arXiv:2404.09656*.

Zihui Gu, Xingwu Sun, Fengzong Lian, Zhanhui Kang, Chengzhong Xu, and Ju Fan. 2024. Dingo: Towards diverse and fine-grained instruction-following evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18108–18116.

Metin N Gurcan, Laura E Boucheron, Ali Can, Anant Madabhushi, Nasir M Rajpoot, and Bulent Yener. 2009. Histopathological image analysis: A review. *IEEE reviews in biomedical engineering*, 2:147–171.

Sunan He, Yuxiang Nie, Zhixuan Chen, Zhiyuan Cai, Hongmei Wang, Shu Yang, and Hao Chen. 2024. Meddr: Diagnosis-guided bootstrapping for large-scale medical vision-language learning. *arXiv preprint arXiv:2404.15127*.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. 2024. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, 36.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024b. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, Lingpeng Kong, and Qi Liu. 2024c. Vlfeedback: A large-scale ai feedback dataset for large vision-language models alignment. *arXiv preprint arXiv:2410.09421*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Mianxin Liu, Jinru Ding, Jie Xu, Weiguo Hu, Xiaoyang Li, Lifeng Zhu, Zhian Bai, Xiaoming Shi, Benyou Wang, Haitao Song, et al. 2024c. Medbench: A comprehensive, standardized, and reliable benchmarking system for evaluating chinese medical large language models. *arXiv preprint arXiv:2407.10990*.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Kenji Ikamura, Georg Gerber, Ivy Liang, Long Phi Le, Tong Ding, Anil V Parwani, et al. 2023. A foundational multimodal vision language ai assistant for human pathology. *arXiv preprint arXiv:2312.07814*.

Saketh Mattupalli. 2025. Medical named entity recognition (ner).

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

JD Pallua, A Brunner, B Zelger, M Schirmer, and J Haybaeck. 2020. The future of pathology is digital. *Pathology-Research and Practice*, 216(9):153040.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Mehmet Saygin Seyfioglu, Wisdom O Ikezogwo, Fatemeh Ghezloo, Ranjay Krishna, and Linda Shapiro. 2024. Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13183–13192.

Guohao Sun, Can Qin, Huazhu Fu, Linwei Wang, and Zhiqiang Tao. 2024. Self-training large language and vision assistant for medical question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20052–20060.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. GitHub repository.

Jeroen Van der Laak, Geert Litjens, and Francesco Ciompi. 2021. Deep learning in histopathology: the path to the clinic. *Nature medicine*, 27(5):775–784.

Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. 2022. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical image analysis*, 81:102559.

Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. 2024. Mmed-rag: Versatile multimodal rag system for medical vision language models. *arXiv preprint arXiv:2410.13085*.

Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, et al. 2024a. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. *arXiv preprint arXiv:2408.02900*.

Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. 2024b. V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. *arXiv preprint arXiv:2411.02712*.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. 2024. Llava-critic: Learning to evaluate multimodal models. *arXiv preprint arXiv:2410.02712*.

Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Xiaoshui Huang, Zhiyong Wang, Lu Sheng, Lei Bai, et al. 2024. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Advances in Neural Information Processing Systems*, 36.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2024. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.

Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, et al. 2024. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, pages 1–13.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2023. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Kangyu Zhu, Peng Xia, Yun Li, Hongtu Zhu, Sheng Wang, and Huaxiu Yao. 2024a. Mmedpo: Aligning medical vision-language models with clinical-aware multimodal preference optimization. *arXiv preprint arXiv:2412.06141*.

Linzhen Zhu, Haoting Zhang, Haibo Gu, and Jianying Zhou. 2024b. The pathology biopsy represents the “gold standard” for diagnosis: a case report. *Diagnostic Microbiology and Infectious Disease*, 108(2):116138.

Appendix for EAGLE

Abstract. In this appendix, we provide additional information about the proposed EAGLE. Appendix A illustrates additional experiments of the proposed EAGLE with human evaluation. Appendix B shows the extra qualitative result on Quilt-VQA. Appendix C provides more results on instruction alignment. Appendix D demonstrates the prompts for self-preference creation. Appendix E discusses the reliability of rejected samples.

A Human Evaluation

In this section, we evaluate the performance by a human evaluation involving 5 users, each was asked to rank 5 LVLMs among 84 cases. This study assesses whether the answer meets the instructions and complies with the doctor's behavior. The ranking was awarded 5 points for first place, which decreased to 1 point for fifth place. As shown in Table 8, EAGLE emerged as the top performer with an average score of 3.76, followed by Quilt-LLaVA (3.20), LLaVA-Med (2.90), LLaVA (2.67), and STLLaVA-Med (2.46). Specifically, in Fig. 6, EAGLE showed the best overall performance, ranking first 193 times, significantly better than other models. The Quilt-LLaVA model performed relatively evenly, ranking first 105 times and second 133 times, indicating that the model can maintain a good performance level in most cases. However, it still ranked fifth 98 times, indicating that its performance is somewhat unstable. LLaVA-Med and LLaVA performed poorly overall, with a higher frequency of ranking at the bottom.

Table 8: Average score for human evaluation.

EAGLE	Quilt-LLaVA	LLaVA-Med	LLaVA	STLLaVA-Med
3.76	3.20	2.90	2.67	2.46

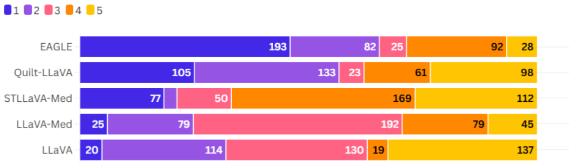


Figure 6: Frequency of occurrence of different rankings.

B Example comparison

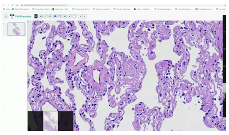
Visual question answering Table 9 shows the comparative performance of various LVLMs in histopathological VQA. It shows a confusing case of Nonalcoholic Steatohepatitis (NSIP) to verify the performance of different models. The base LLaVA model provided a rudimentary identification of key features, namely inflammation and fibrosis, though lacking in diagnostic depth. LLaVA-Med, while acknowledging the diagnostic challenges, emphasized the absence of specific histological features, potentially reflecting a more cautious clinical approach. Notably, the response of STLLaVA-Med deviated significantly from professional medical discourse, inappropriately incorporating humor into its interpretation. In contrast, Quilt-LLaVA exhibited superior performance by accurately describing the pathological pattern and distinguishing features, including the characteristic 'patchy' appearance and interstitial inflammation patterns. EAGLE demonstrated the most comprehensive analysis, providing detailed observations of uniform interstitial patterns while correctly noting the absence of architectural distortion and honeycomb features.

C More Results on Instruction Alignment

VQA with visual prompting Table 11 shows cases on the histopathology VQA with visual prompting on different LVLMs. The models demonstrated varying levels of descriptive depth and pathophysiological understanding when presented with visual prompting questions focusing on specific regions highlighted by red circles. The base LLaVA model provided fundamental observations, identifying bile ducts and morphological changes, though lacking detailed cellular characterization. LLaVA-Med and STLLaVA-Med offered similar basic descriptions of bile presence, with STLLaVA-Med specifically noting bile pigment accumulation. Notably, Quilt-LLaVA demonstrated superior analytical capability by providing a comprehensive description that included specific cellular changes (swollen hepatocytes with granular cytoplasm), pathophysiological mechanisms (impaired bile flow), and the ability to infer cholestasis from observed features. EAGLE contributed additional clinical context by linking the histological findings to potential etiologies, particularly drug and toxin-related causes while maintaining accurate morphological descriptions.

Instruction Alignment Table 12 shows the comparison of different LVLMs on instruction alignment. The evaluation was implemented to assess

Table 9: Example comparison of histopathology visual question answering on different LVLMs. Green for correct and red for wrong.

Input example, Histopathology Image and Question:	
User	 What characterizes the cases of NSIP?
LLaVA	The cases of NSIP (Nonalcoholic Steatohepatitis) are characterized by the presence of inflammation and fibrosis in the liver.
LLaVA-Med	The cases of NSIP (nonspecific interstitial pneumonia) are characterized by a lack of specific histological features. This makes it difficult to accurately diagnose NSIP based solely on histological examination.
STLLaVA-Med	The image shows a histopathy of NSIP, which stands for ḧistopathy of normal pulmononium.Ït is important to note that this is a play on words, not a medical condition. The image is meant to be humorous and should not be taken literally.
Quilt-LLaVA	In the cases of Non-Specific Interstitial Pneumonia (NSIP), there are several key features to note. NSIP is characterized by a uniform or 'non-patchy' appearance, which is a distinguishing factor from other interstitial lung diseases. This uniformity is seen in the image as a consistent pattern of interstitial inflammation and fibrosis.
EAGLE	The histopathological features of NSIP can vary widely, but they typically include a uniform pattern of interstitial inflammation and fibrosis, with no significant architectural distortion or honeycombing. The fibrosis in NSIP is usually uniform and lacks the characteristic features of usual interstitial pneumonia (UIP), such as a patchy distribution and fibroblastic foci.
GT	NSIP cases are characterized by uniform expansion of the alveolar septa. It's a very homogeneous process in contrast to cases of UIP.

the performance of various LVLMs against GPT-generated responses in histopathological image interpretation. During the evaluation process, each LVLM will receive two scores (model score / GPT score) and the evaluation basis will be given. Significant variations in diagnostic reasoning and response quality were observed in this comparative analysis of LVLMs addressing histopathological assessment of giant cell presence. The evaluation focused on the models' ability to interpret the absence of giant cells and its diagnostic implications. LLaVA and LLaVA-Med provided basic interpretations but demonstrated limited clinical reasoning, with scores of 4/8 reflecting their inaccurate and incomplete analysis. STLLaVA-Med showed improved understanding by discussing fibroblastic implications, though lacks relevance to the diagnosis. A notable advancement was observed in Quilt-LLaVA's response (7/8), which demonstrated sophisticated clinical reasoning by incorporating cellular architecture considerations and broader histological context. However, the explanation was less focused and introduced additional considerations. EAGLE achieved the highest score (9/8), providing the most comprehensive analysis by integrating multiple diagnostic aspects, including abnormal cell shapes, tissue arrangements, and the importance of whole-slide examination. The model particularly excelled in emphasizing the need for comprehensive tissue examination and additional clinical correlation.

D Self-Preference Creation

The prompt of the expert-verified chosen sampling is carefully crafted by the pathological expert. The prompt $\mathbf{P_c}$ is structured as: `You are an AI assistant who specializes in pathological diagnosis questions and answers. Please generate a high-quality answer to the questions.`

The prompt $\mathbf{P_r}$ of the expert-guided rejected sampling is defined as: `Please generate a low-quality-answer to the question, that is highly relevant but not semantically identical to the questions above from the user.`

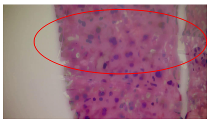
E Reliability of rejected samples

It is important to note that the preference learning framework is designed to leverage the relative quality differences between the "chosen" and "re-

Table 10: Performance comparison of LVLM instruction alignment abilities, measured by the relative score via GPT-4 evaluation. Question type mainly consists of conversation and detailed description. Sub-domain contains Bone, Breast, Cytopathology and Hepatopathology (Cyto), Dermatopathology (Derm), Endocrine (Endo), Gastrointestinal (Gastro), Genitourinary (Geni), Gynecologic (Gyne), Head and Neck (H& N), Hematopathology (Hema), Neuropathology (Neuro), Pulmonary (Pulm), Renal, and Soft tissue (Soft). The number of each subdomain is shown in brackets.

LVLMs	**Question Types**		**Sub-Domains**														**Overall**
	Conv	Desc	Bone	Breast	Cyto	Derm	Endo	Gastro	Geni	Gyne	H&N	Hema	Neuro	Pulm	Renal	Soft	
	(256)	(70)	(25)	(23)	(23)	(21)	(23)	(23)	(23)	(22)	(24)	(22)	(24)	(25)	(23)	(25)	(326)
LLaVA (Liu et al., 2024b)	55.0	40.0	51.5	55.2	44.8	48.2	54.4	47.9	50.9	58.3	56.4	47.2	59.4	45.1	48.8	54.2	51.5
LLaVA-Med (Li et al., 2024b)	61.8	42.9	58.5	63.5	51.7	66.2	46.2	50.3	61.2	57.8	64.1	55.6	58.2	53.2	59.8	61.0	57.4
STLLaVA-Med (Sun et al., 2024)	69.1	47.9	69.4	67.1	50.9	72.8	62.2	63.9	67.9	65.8	73.2	56.5	63.5	53.7	65.9	69.4	64.2
Quilt-LLaVA (Seyfioglu et al., 2024)	72.0	58.4	77.4	75.1	59.8	**75.4**	**79.9**	59.3	**78.4**	61.0	70.4	**64.9**	61.7	56.4	**66.5**	80.5	68.8
EAGLE	**72.7**	**58.8**	**81.0**	**75.8**	**60.3**	66.9	66.3	**59.7**	73.9	**76.1**	**75.6**	64.3	**67.9**	**59.6**	65.9	**81.0**	**69.5**

Table 11: Example comparison of histopathology VQA with visual prompting on different LVLMs. Green for correct and red for wrong.

Input example, Histopathology Image and Question:	
User	 What is the pattern of cholestatic hepatitis associated with ? Tip: Focus on image areas highlighted by the red circle(s).
LLaVA	The pattern of cholestatic hepatitis associated with the red circle(s) in the image is characterized by the presence of numerous small bumps or bile ducts.
LLaVA-Med	The pattern of cholestatic hepatitis associated with is characterized by the presence of bile ducts, which are marked by the red circle(s) in the image.
STLLaVA-Med	The pattern of cholestatic hepatitis in this image is associated with the presence of bile pigment granules in the portal tracts. These granules are visible in the image areas highlighted by the red circle(s).
Quilt-LLaVA	The highlighted areas in the image, as indicated by the red circle(s), show a pattern of cholestatic hepatitis. This condition is characterized by the accumulation of bile within the liver due to impaired bile flow. The hepatocytes in these areas may appear swollen and have a pale, granular cytoplasm. The presence of cholestasis can be inferred from the presence of these features.
EAGLE	The image shows a pattern of cholestatic hepatitis, which is characterized by the presence of bile within the hepatocytes. This may be related to drugs and drug toxins. The pattern of cholestatic hepatitis is typically associated with a variety of liver diseases.
Ground-Truth	The pattern of cholestatic hepatitis seen in the image is associated with drug and medication toxins, as well as hepatitis A and E, as these can produce cholestatic injury features.

jected” responses, rather than relying solely on the absolute correctness of the “chosen” samples. Direct Preference Optimization (DPO), in particular, demonstrates strong robustness to noise and label inconsistencies through several key properties:

Noise Samples Have Local Impact: DPO only uses pairwise preference comparisons, so a mislabeled pair affects only the loss from that specific comparison. It does not distort the absolute scores or global preference structure. The loss function for a noisy pair is given as:

$$\mathcal{L}^{\text{noisy}} = -\log \sigma \Big(-\beta \big(\log \pi_\theta(y^w|x) - \log \pi_\theta(y^l|x)\big)\Big) \tag{5}$$

where the flipped gradient only affects the local direction for the specific triple (x, y^w, y^l). The rest of the dataset remains unaffected.

Correct Preferences Outweigh Noisy Ones: Since DPO relies on many preference pairs, noisy labels are diluted by the majority of correct comparisons. The model’s update direction reflects the dominant preference signal. The update direction for the model parameters is computed as:

$$\nabla_\theta \mathcal{L}_{\text{DPO}} = -\beta \left(\sigma(\beta\Delta) - 1\right) \nabla_\theta \Delta$$

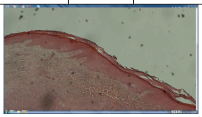
$$\text{where} \quad \Delta = \log \pi_\theta(y^w|x) - \log \pi_\theta(y^l|x) \tag{6}$$

Even if some $\Delta < 0$ due to noise, most updates push in the correct direction when $\Delta > 0$ dominates.

DPO Is Inherently Robust to Label Scale Bias: Unlike reward modeling or pointwise RLHF methods, DPO does not require consistent or absolute reward values. It only compares preference direction via log-ratios:

$$\Delta R(x) \propto \log \pi_\theta(y^w|x) - \log \pi_\theta(y^l|x) \tag{7}$$

Table 12: Example comparison of histopathology VQA with GPT evaluation. Each LVLM will receive two scores (model score / GPT score). GPT's analysis of generated answers is highlighted.

Input example, Histopathology Image and Question:			
	Gen .Answer	Score	Analysis
User	 How common is it to observe acanthosis in large tumors, and what does its presence signify?		
LLaVA	The absence of giant cells in the area of the image implies that the diagnosis might be different from what was initially suspected. Giant cells are often associated with certain diseases, such as giant cell arteritis or temporal arteritis, which are inflammatory conditions affecting the blood vessels. However, the absence of giant cells in this area suggests that the diagnosis might be more accurate or that the pathologist has identified a different condition. It is essential to consider other factors, such as the patient's symptoms, medical history, and additional diagnostic tests, to confirm the correct diagnosis.	4/8	Assistant 2, on the other hand, provides an **inaccurate and less relevant response**. It incorrectly suggests that the absence of giant cells might imply a different diagnosis, such as giant cell arteritis or temporal arteritis, which are unrelated to giant cell tumors of bone. This response **lacks accuracy and relevance to the context of the question**, which is specifically about giant cell tumors of bone. Additionally, it introduces unrelated conditions, which could confuse the reader. Overall, Assistant 1's response is more aligned with the context and question, leading to a higher score.
LLaVA-Med	The absence of giant cells in this area suggests that the diagnosis of giant cell tumor (GCT) may not be accurate. GCT is a type of bone tumor characterized by the presence of giant cells. If the diagnosis is incorrect, it may lead to inappropriate treatment and management of the patient's condition.	4/8	Assistant 2, on the other hand, **incorrectly suggests** that the absence of giant cells in the area might imply an inaccurate diagnosis of a giant cell tumor. **This is misleading because it does not take into account the complexity of histopathological diagnosis**, where the presence or absence of certain features in isolated areas does not solely determine the diagnosis. This response lacks accuracy and could lead to misunderstanding, thus receiving a lower score.
STLLaVA-Med	The absence of giant cells in the histopathological image suggests that the lesion is likely a benign fibroblastic tumor. Fibroblastic tumors are typically slow-growing, noncancerous growths that arise from fibroblasts, which are cells that produce collagen and other extracellular matrix components. The absence of giant cells in this area indicates that the tumor is not malignant and may not be aggressively invading the surrounding tissue. However, it is important to consult a healthcare professional for a thorough evaluation and proper diagnosis of the specific type of fibroblastic tumor and to determine the appropriate course of treatment.	4/8	Assistant 2, on the other hand, **incorrectly suggests** that the absence of giant cells implies the lesion is likely a benign fibroblastic tumor. This is misleading because the absence of giant cells in one area does not automatically indicate a different type of tumor. **The response lacks accuracy and relevance to the context of diagnosing a giant cell tumor of bone.** Additionally, **it introduces unnecessary information about fibroblastic tumors, which is not directly related to the question**. Therefore, Assistant 2's response is less helpful and less accurate, resulting in a lower score.
Quilt-LLaVA	The absence of giant cells in this area is indeed intriguing. Giant cells are often seen in certain types of tumors, including giant cell tumors, which are typically benign. However, their absence does not necessarily rule out a benign diagnosis. It's important to consider the overall cellular architecture and other histological features. Hint: Consider the presence of other cell types in the area. For instance, are there any spindle-shaped cells or any signs of hemorrhage? Also, think about the patient's age and symptoms. How might these factors influence your diagnosis?	7/8	Assistant 2 also offered a relevant response, noting that the absence of giant cells does not rule out a diagnosis. **However, the explanation was less focused and introduced additional considerations**, such as spindle-shaped cells and patient factors, which, while potentially useful, were not directly related to the user's question about the absence of giant cells. This made the response slightly less precise and potentially more confusing for the user. The level of detail was higher, but not all of it was directly relevant to the question asked.
EAGLE	The absence of giant cells in this area of the histopathology image is indeed intriguing. Giant cells are often a key feature in certain types of tumors, such as giant cell tumors. However, their absence does not necessarily rule out the possibility of a tumor. It's important to remember that the presence or absence of giant cells is just one aspect of the overall histopathological picture. In the context of a tumor, the presence of other cellular features such as abnormal cell shapes, sizes, and arrangements, as well as the presence of mitotic figures, can be indicative of malignancy. Additionally, the location of the tumor and its relationship with surrounding structures can provide valuable clues. In this case, the absence of giant cells in the area of focus doesn't necessarily exclude the possibility of a tumor. It's possible that giant cells are present in other areas of the tissue not visible in this particular image. Therefore, a comprehensive examination of the whole slide, including multiple areas of the tissue, would be necessary to make a definitive diagnosis. In conclusion, while the absence of giant cells in this area is noteworthy, it doesn't provide a definitive diagnosis. Further examination of the whole slide, along with clinical correlation and possibly additional tests, would be necessary to confirm the diagnosis.	9/8	Assistant 2 offers a **more detailed response**, explaining not only the significance of the absence of giant cells but also discussing other histopathological features that might be considered in the diagnosis. It provides a more comprehensive view by mentioning the importance of examining the whole slide and correlating with clinical information. **This response is more informative and better addresses the complexity of histopathological diagnosis, making it more helpful and relevant to the user's question.**

Hence, DPO is invariant to reward scaling or additive shifts in reward annotations. This makes it more robust to subjective or inconsistent human labels.

To further perform a quantization analysis, we track the reward progression of “chosen” and “rejected” samples during training. As shown in Table 13, as training progresses, the reward for chosen samples increases, while the reward for rejected samples decreases significantly. This results in a progressively larger delta between the two. The growing gap correlates with improved accuracy from 53% to 89%. These findings suggest that the reliability of rejected samples has minimal impact on preference training, as the focus is on the relative quality difference.

Table 13: Reward progression and accuracy improvement during training

Iteration	**10**	**20**	**50**	**100**
Reward of Chosen Sample	0.0020	0.0025	0.0267	0.0321
Reward of Rejected Sample	-0.0160	-0.1911	-2.0880	-6.7071
Delta	0.0180	0.1936	2.1148	6.7392
Accuracy	**0.5312**	**0.8125**	**0.8281**	**0.8906**