Walking the Tightrope: Autonomous Disentangling Beneficial and Detrimental Drifts in Non-Stationary Custom-Tuning

Xiaoyu Yang, Jie Lu, En Yu
Australian Artificial Intelligence Institute (AAII),
Faulty of Engineering and Information Technology,
University of Technology Sydney, Australia.
xiaoyu.yang-3@student.uts.edu.au; {jie.lu,en.yu-1}@uts.edu.au

Abstract

This paper uncovers a critical yet overlooked phenomenon in multi-modal large language models (MLLMs), especially for chest diagnosis: detrimental concept drift within chain-of-thought (CoT) reasoning during non-stationary reinforcement fine-tuning (RFT), where reasoning token distributions evolve unpredictably, thereby introducing significant biases in final predictions. To address this, we are pioneers in establishing the theoretical bridge between concept drift theory and RFT processes by formalizing CoT's autoregressive token streams as non-stationary distributions undergoing arbitrary temporal shifts. Leveraging this framework, we propose a novel autonomous counterfact-aware RFT that systematically decouples beneficial distribution adaptation from harmful concept drift through concept graph-empowered LLM experts generating counterfactual reasoning trajectories. Our solution, Counterfactual Preference Optimization (CPO), enables autonomous and stable RFT in non-stationary environments, particularly within the medical domain, through custom-tuning of counterfactual-aware preference alignment. Extensive experiments demonstrate our superior performance of robustness, generalization and coordination within RFT. Besides, we also contribute a large-scale dataset CXR-CounterFact (CCF), comprising 320,416 meticulously curated counterfactual reasoning trajectories derived from MIMIC-CXR. Our code and data are public at: https://github.com/XiaoyuYoung/CPO.

1 Introduction

Reinforcement Fine-Tuning (RFT) [1, 2] has emerged as a promising paradigm for domain-specific customization of multi-modal large language models (MLLMs) [3–5], demonstrating remarkable capability in facilitating efficient domain shift with minimal data requirements, particularly for medical downstream tasks. However, the reinforcement-driven custom-tuning is fundamentally challenged by non-stationary environmental dynamics, especially for inherent domain-specific data characteristics such as long-tailed distributions in medical diagnosis, and systemic data imperfections including noise and sparsity. This complex synergy induces latent concept drift that progressively disaligns the model's representation space from domain reality, culminating in catastrophic error propagation that particularly jeopardizes the reliability of MLLMs in safety-critical applications like radiology report generation.

Concept drift theory [6, 7] provides a new perspective for analyzing the domain shift of RFT in non-stationary custom-tuning, which focuses on the unpredictable distribution changes in data streams.

Correspondence to Jie Lu and En Yu

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

We posit that the autoregressive decoding paradigm inherent to MLLMs can be characterized as a sequential token stream generation process. Within this framework, each token generation step propagates through the model's internal reasoning pathways, which remain opaque to external observation, while manifesting inherent stochasticity in the evolving token probability distributions across successive decoding iterations.

Within the concept drift framework, our analysis reveals critical limitations in reinforcement fine-tuning approaches that depend on verifiable rewards in chain-of-thought (CoT) [2]:

Observation 1.1. *Specifically, while RFT operate through optimal reasoning pathway selection to maximize outcome certainty, we empirically observe that MLLM-generated CoT processes in specialized domains frequently demonstrate susceptibility to concept drift. This progressive deviation in intermediate reasoning ultimately induces substantial output divergence in non-stationary environments.*

Intuitively, we provide a representative case study that demonstrates this phenomenon in clinical reasoning contexts as presented in Fig.1. When diagnosing chest DR images, the model generates a reasoning trajectory containing the statement: "Asymmetric lung opacity in the right middle lobe is concerning for pneumonia." We found that in the token-level probability, "lung opacity" shows negligible differentiation from its ambiguous counterpart "opacity". Despite this minimal probabilistic disparity in the thinking process, our diagnostic outcome distribution analysis presents a radical divergence in final predictions as exhibited in Fig.1. In particular, the diagnosis is completely opposite in terms of atelectasis, cardiomegaly and pneumonia.

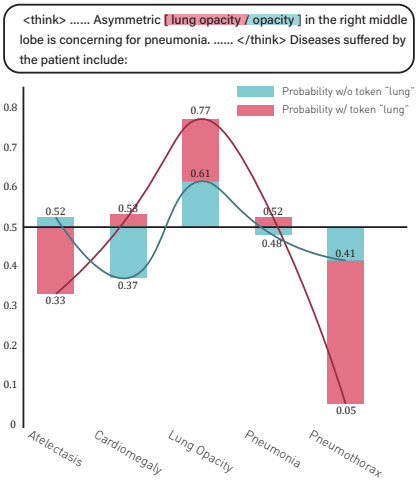


Figure 1: **Concept Drift in RFT's reasoning for chest diagnosis.** Despite analogous occurrence probabilities of "lung opacity" (in red) and "opacity" (in blue) tokens during the CoT, non-stationarity induces significant bad distributional drift in clinical conclusions, especially the opposite diagnosis of atelectasis, cardiomegaly and pneumonia.

Therefore, summarizing the above challenges of reinforced fine-tuning in MLLMs, it raises the important question:

How to adapt thinking process to concept drift under non-stationary reinforced custom-tuning?

Inspired by causal inference [8–10], we develop Counterfactual Preference Optimization (CPO), a principled approach that systematically perturbs reasoning trajectories to discriminate between beneficial distribution adaptation and detrimental concept drift.

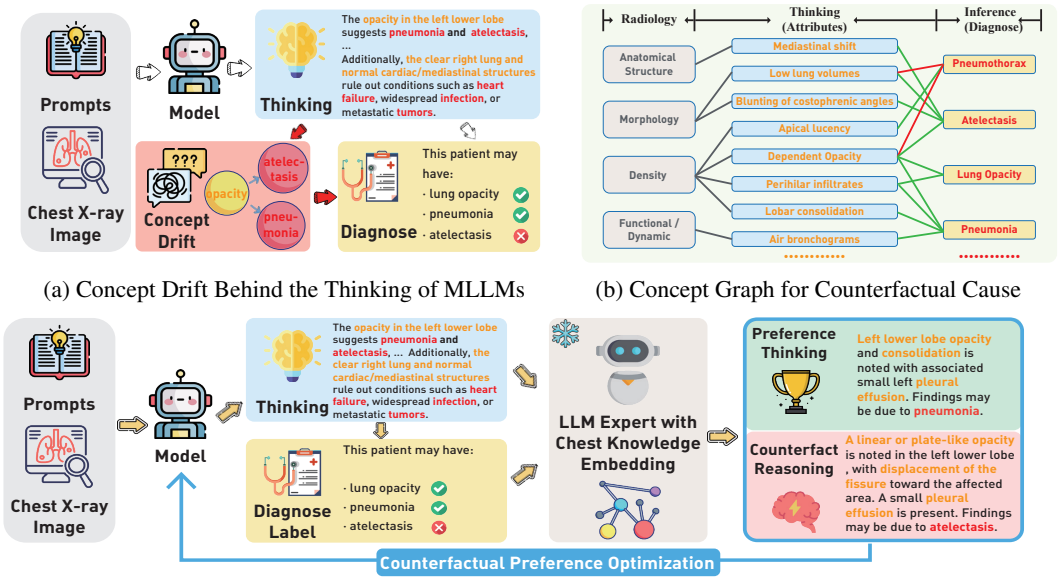
Firstly, we construct a hierarchical concept graph that codifies domain-specific knowledge structures through triadic relation embeddings, including positive correlation, irrelevance, and opposition. Subsequently, we structurally embed the hierarchically structured concept graph into the LLM's reasoning architecture as an expert-guided module, automatically generating semantically-constrained counterfactual inference paths. Consequently, during reinforced custom-tuning, we formulate a dual-preference optimization objective that jointly maximizes likelihood alignment with human preferences while minimizing similarity to adversarially generated counterfactual paths, thus achieving decoupling of beneficial domain adaptation from detrimental concept drift. Finally, we contribute CXR-CounterFact (CCF), the chest diagnosis preference dataset comprising 320,416 fine-curated counterfactual reasoning trajectories derived from MIMIC-CXR [11] radiologic findings, aiming to validate our method and catalyze research advancements in counterfactual-aware reinforcement fine-tuning paradigms

In summary, our paper mainly makes the following contributions:

1. First, we establish a novel theoretical framework that formalizes autoregressive token generation in MLLMs through the lens of concept drift theory, enabling systematic identification and causal analysis of detrimental reasoning divergence during non-stationary reinforced custom-tuning.

2. Second, we propose Counterfactual Preference Optimization (CPO), which synergises structured domain-specific knowledge with systematic counterfactual intervention, driving the MLLMs with preference-aligned reinforcement learning. By embedding learnable concept graphs as the expert and generating adversarially-constrained reasoning trajectories, our approach achieves substantial decoupling between beneficial distribution adaptation and detrimental concept drift.
3. Third, we conduct comprehensive empirical validation across various clinical benchmarks for chest radiograph, including disease classification, diagnostic report generation and zero-shot generalization. The superior results demonstrate statistically significant improvements in robustness, generalization, and accuracy of our method under non-stationary custom-tuning. Besides, we also provide ablation experiments to validate the effectiveness of various modules.
4. As a pioneer contribution to the community, we introduce CXR-CounterFact (CCF), a large-scale dataset comprising 320,416 meticulously curated counterfactual reasoning trajectories derived from MIMIC-CXR.

2 Methodology



(a) Concept Drift Behind the Thinking of MLLMs

(b) Concept Graph for Counterfactual Cause

(c) Counterfactual Preference Optimization

Figure 2: The main contributions of our methods. (a) By formalizing autoregressive CoT generation as a stream of next-token prediction actions under the theoretical lens of concept drift, we reveal that even minor perturbations in reinforced fine-tuning can induce unpredictable distributional changes of final predicted results. (b) To disentangle detrimental drift, we introduce the concept graph that generates radiologically plausible counterfactual CoTs through controlled attribute perturbations. Green lines represent attributes that are positively correlated with the disease, while red denote they are exclusive. (c) We propose counterfactual preference optimization to drive the reinforced custom-tuning of MLLMs, enabling generalized CoT reasoning in non-stationary environments through disentanglement of beneficial domain adaptation from spurious concept drift, thereby achieving robust human-aligned decision-making via preference distillation.

2.1 Underlying Concept Drift Behind Thinking

In this section, we first extend the concept drift theory to reinforced custom-tuning, highlighting the phenomenon wherein the distributional characteristics of targets undergo arbitrary changes over the course of the thinking process. Operating through recursive on-policy sampling, the MLLM π

autoregressively generates the token at position j in the reasoning chain, conditioned on both the visual input image v, the textual prompt l, and the partial token sequence $t_{<j}$ of the CoT trajectory:

$$t_j \sim \pi(\cdot|v, l, t_{<j}) \tag{1}$$

Therefore, we formally define concept drift behind the thinking as follows:

Definition 2.1. *The MLLM's autoregressive reasoning trajectory manifests as a thinking stream $S_{0,i} = \{s_0, ..., s_i\}$, where each cognitive state $s_j = (t_{<j}, z_j)$ encapsulates all tokens generated so far $t_{<j}$ and its latent predicted distribution z_j of the results by $t_{<j}$. Therefore, in position i, $S_{0,i}$ follows a certain distribution $F_{0,i}(x, z)$, thus the concept drift behind the thinking can be formalized as:*

$$\exists i : P_i(t, z) \neq P_{i+1}(t, z) \tag{2}$$

where the joint probability $P_i(t, z)$ can be decomposed as $P_i(t, z) = P_i(t) \times P_i(z|t)$.

Consequently, this concept drift framework behind the thinking of MLLMs enables simultaneous characterization of temporal dynamics in Chain-of-Thought reasoning, formalized as the concept drift process $P_i(t)$, and its induced probabilistic divergence $P_i(z|t)$, capturing the evolving discrepancy between intended and actual outcome distributions throughout cognitive progression.

To adapt the reinforced custom-tuning to concept drift behind the CoT, it is essential to adapt the model to align with the evolving thinking distribution under non-stationary environment, which can be formally defined as:

$$\min_{\pi^{(i)}, \pi^{(i+1)}, ..., \pi^{(i+\tau)}} \sum_{i}^{i+\tau} \mathcal{L}(\prod_{j=1}^{L} \pi^{(i)}(t_j^{(i)}|v, l, t_{<j}^{(i)}), y^{(i)}), \tag{3}$$

where $\prod_{j=1}^{L} \pi^{(i)}(t_j^{(i)}|v, l, t_{<j}^{(i)})$ denotes the probability of CoT token sequence, $y^{(i)}$ represents the preferred CoT, L symbolics the max length of tokens within the CoT, and $\pi^{(i)}$ signifies the MLLM at the cognitive status i. And the model is driven by the target metric $\mathcal{L}$ continuously to adapt the drift in a given time period $[i, i + \tau]$. Thus, we get the optimization object within the concept drift framework.

2.2 Disentangling Concept Drift with Counterfactual Causes

The optimization objective formalized Eq.3 necessitates disentanglement of two competing goals: advantageous policy-induced domain adaptation and versus pathological concept drift arising from suboptimal policy execution, which are both sampled from policy π within the time period $[i, i + \tau]$. However, it is challenging to determine the optimal preferred CoT and explicitly judge which strategies will cause unpredictable changes in tokens.

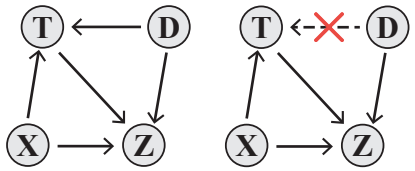


Figure 3: **Structural Causal Graph. X**: Inputs, **Z**: Prediction Results, **T**: Chain-of-Thought, and **D**: Latent Concept Drift within CoT under Non-stationary Reinforced Custom-Tuning.

Fortunately, counterfactual causes provide an explicit manner to decouple these two competing goals. We construct a structural causal graph [8, 9] to formula the causal relationship among elements as discussed in Section 2.1, including inputs (X) consisting of image v and prompt l, prediction result (Z), Chain-of-Thought (T) and the concept drift in the reinforcement custom-tuning (D) as illustrated in Fig.3, where $A \rightarrow B$ denotes that A is the causer of B. The causal graph of $\{X, Z, T, D\}$ presents the following causal connections:

$(X, D) \rightarrow T$: This link denotes the chain-of-thought T derived from the inputs X through policy π is under the impact of latent concept drift D.

$(X, T, D) \rightarrow Z$: This link presents that, apart from the regular reasoning pathway of$(X, T) \rightarrow Z$, the prediction is also impacted by the concept drift D through the pathway of $D \rightarrow T \rightarrow Z$.

In the constructed structural causal model, nodes D and T are formally characterized as the confounder and mediator [12], respectively. The confounding variable D induces bias through the

backdoor path $X \leftarrow D \rightarrow Z$ [8], simultaneously influencing both the mediator T and the outcome variable Z. This interference systematically distorts the estimation of CoT reasoning's causal effect on model predictions, particularly under non-stationary adaptation scenarios. The resulting spurious correlations manifest as concept drift artifacts that propagate through the mediation pathway $X \rightarrow T \rightarrow Z$, ultimately compromising the stability of customized model tuning.

Building upon the above analysis grounded in cause, we formally decouple the concept drift dynamics in Chain-of-Thought reasoning of Eq.3 grounded in the cause. By constructing interventional distributions through **do** operations, $P(Z|\mathbf{do}(T = t), D = d)$, we quantify the latent causal effect:

$$\psi = \mathbb{E}[Z_{T\leftarrow t, D\leftarrow d} - Z_{T\leftarrow t', D\leftarrow d}] \tag{4}$$

where the potential outcome $T \leftarrow t$ represents the counterfactual scenario when forcibly maintaining the mediator T at value t, while preserving the confounder state d. This formulation explicitly isolates the front-door effect $X \rightarrow T \rightarrow Z$ from backdoor concept drift propagation $X \leftarrow D \rightarrow Z$.

2.3 Embedding Counterfactual Causes with LLM expert

Having operationalized concept drift decoupling through controlled counterfactual interventions in Section 2.2, we identify the generation of autonomous counterfactual causes as the subsequent bottleneck that requires maintaining causal consistency within the chain-of-thought while avoiding semantic entanglement.

Accordingly, inspired by [13, 14], we constructed a hierarchical concept graph for custom-tuning through autonomous knowledge extraction from chest radiograph reports. It systematically organizes medical concepts into four semantic dimensions: disease entities, radiographic features, clinical relationships, and taxonomies. Specifically, the knowledge extraction pipeline leverages a medical domain-adapted large language model with radiological prior knowledge, to process 160,208 chest X-ray reports from the MIMIC-CXR dataset [11]. Through iterative semantic parsing, the model autonomously identifies 12 distinct pulmonary disease entities accompanied by 53 clinically relevant attributes. These attributes are meticulously annotated across four diagnostic categories, namely morphological alterations, density anomalies, anatomical deviations, and functional/dynamic indicators. To capture clinical interdependencies, we formalize three types of ontological relationships, including association, irrelevance, and exclusion. For instance, the framework automatically detects pathophysiological contradictions between emphysema-associated hyperinflation and atelectasis-related lung volume reduction, leading to the exclusion relationship between emphysema and atelectasis. As a toy example illustrated in Fig.2b, the resulting concept graph provides multi-relational representations where each disease entity is instantiated with its associated attributes and constraint relationships, enabling structured reasoning about pulmonary pathology.

Through systematic integration of the concept graph, medical-customized LLM evolves as an expert with causal prior knowledge of chest radiology, which effectively simulates radiologists' differential diagnosis protocols, and synthesises counterfactual diagnostic narratives through controlled feature perturbation while maintaining radiological plausibility, as exhibited below:

2.4 Reinforced Custom-Tuning with Counterfactual Thinking

Obtaining counterfactual diagnosis in Section.2.3, we propose Counterfactual Preference Optimzation (CPO) to drive the reinforced custom-tuning of the multi-modal large language models.

Formally, we have decomposed the CoT generation process into a stream of next token prediction actions $S_{0,i} = \{s_0, ..., s_i\}$, where each cognitive state $s_j = (t_{<j}, z_j)$ encapsulates all tokens generated so far $t_{<j}$ and its latent predicted distribution z_j of the results by $t_{<j}$, as exhibited in Definition 2.1. At timestep j, the action t_j is sampled from the policy $\pi(\cdot|v, l, t_{<j})$ where t_j can be any token in the vocabulary. After each action, the resulting state s_{j+1} is the concatenation of the current state s_j and the action t_j with its latent predicted results:

$$s_{j+1} = (t_{<j} \circ t_j, P_j(z|t_{<j} \circ t_j)), 0 \leq j \leq L \tag{5}$$

where $\circ$ denotes the concatenation between tokens stream $t_{<j}$ and action token t_j, L represents the maximum length of CoT, and P is the latent predicted distribution of results derived by $t_{<j}$ as presented in Eq.2. As the start of the chain-of-thought, a_0 is usually the token <think>. While it produces the </think> token, the resulting state s_{L+1} is the terminal state, thereby concluding one

Ground Truth
Findings: Moderate cardiomegaly is increased. No focal consolidation or pneumothorax. There is a slight blunting of the costophrenic angles, which may indicate small pleural effusion or scarring. There is increased density at the perihilar regions which may indicate pulmonary vascular congestion.
Diagnosis: The disease of this patient is Cardiomegaly.
Generated Counterfactual Reasoning of Pneumonia.
Findings: PA and lateral views of the chest show moderate cardiomegaly. Focal consolidation is noted in the right lower lobe with accompanying bronchial airspace opacification. No pneumothorax is observed. Slight blunting of the costophrenic angles suggests the presence of a small pleural effusion. Increased density at the perihilar regions indicates pulmonary vascular congestion, but also suggests possible pneumonia.
Diagnosis: The disease of this patient is Pneumonia.

Table 1: Example to illustrate the generated counterfactual diagnosis. Ground Truth denotes the original diagnosis report from MIMIC-CXR, and the bottom is the counterfactual report designed for pneumonia. Underline indicates generated counterfactual diagnostic attributes through controlled perturbation while maintaining radiological plausibility, leading to the counterfactual diagnosis.

chain-of-thought generation process. We regard the report issued by the professional doctors as the chain-of-thought preferred by humans, which are positive samples. As for negative samples, we generate counterfactual CoT for diagnostic report instances as our negative samples, that is, perturbing specific radiological features to interfere with the diagnosis results according to our proposed concept graph. Thereby, the chain-of-thought preferred by humans is considered to be t^+, namely the diagnosis report stemming from the radiologist, while the generated counterfactual CoT is represented by t^-.

Consequently, following the DPO [15], we derive the optimal policy that maximizes the reward function through:

$$\pi_\theta(t|v,l) \propto \pi_{\text{ref}}(t|v,l) \exp\left(\frac{r(v,l,t)}{\beta}\right) \tag{6}$$

where β is a parameter controlling the deviation from the base reference policy π_{ref}, namely the initial supervised fine-tuned (SFT) model, and π_θ denotes the fine-tuning model. With counterfactual effect in Eq.4, the reward difference between human-preferred positive samples and counterfactual samples can be defined as:

$$r(v,l,t^+) - r(v,l,t^-) = \beta \left[\log \frac{\pi_\theta(t^+|v,l)}{\pi_{\text{ref}}(t^+|v,l)} - \log \frac{\pi_\theta(t^-|v,l)}{\pi_{\text{ref}}(t^-|v,l)}\right] \tag{7}$$

Thus, based on the Bradley-Terry model, the counterfactual preference optimization (CPO) is driving the reinforced custom-tuning of the MLLMs through the maximum likelihood objective:

$$\mathcal{L}_{\text{CPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(v,l,t^+,t^-)} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(t^+|v,l)}{\pi_{\text{ref}}(t^+|v,l)} - \beta \log \frac{\pi_\theta(t^-|v,l)}{\pi_{\text{ref}}(t^-|v,l)}\right)\right] \tag{8}$$

In this context, it culminates in counterfactual reinforced custom-tuning, an adaptive framework that effectively differentiates between advantageous domain adaptation and harmful concept drift in non-stationary environments, achieving equilibrium preservation through causal intervention and dynamic policy reinforcement, walking the tightrope.

2.5 Building CXR-CounterFact Dataset for Clinical Reasoning Chains

Since we are pioneers in introducing counterfactual cause into reinforced custom-tuning of MLLMs, we are deeply aware of the scarcity of counterfactual CoT in downstream tasks, especially in the highly professional medical field. Thus, our aspiration is for the model to adeptly acclimate to the concept drift by itself, acquiring abundant knowledge with more and more data, but not exhibiting bias.

In this context, a more realistic training dataset for multi-modal large language models is required to validate their potential to be trained under the non-stationary reinforced custom-tuning. Recognizing the demand for higher-quality multi-modal data with CoT, we develop a datasets called CXR-CounterFact Dataset (CCF), extending the MIMIC-CXR[11] with counterfactual chain-of-thought. This novel dataset introduces 320,416 meticulously curated counterfactual pairs spanning 14 thoracic pathologies, establishing a pioneering large-scale benchmark for causal interpretation in clinical chest X-ray analysis. More details are given in Appendix B.

3 Experiments

In this section, we verify the robustness, generalization and coordination of our proposed counterfactual preference optimization in reinforced custom-tuning in non-stationary environments.

MIMIC-CXR[11] is utilized to train the MLLMs via reinforced custom-tuning for domain adaptation, which presents 371,920 chest X-rays associated with 227,943 imaging studies from 65,079 patients. And images are provided with 14 labels with corresponding free-text radiology reports, namely Atelectasis (Ate.), Cardiomegaly (Car.), Consolidation (Con.), Edema (Ede.), Enlarged Cardiomediastinum (ECM), Fracture (Fra.), Lung Lesion (LL), Lung Opacity (LO), Pleural Effusion (PE), Pneumonia (Pna.), Pneumothorax (Pnx.), Pleural Other (PO), Support Devices (SD) and No Finding (NF).

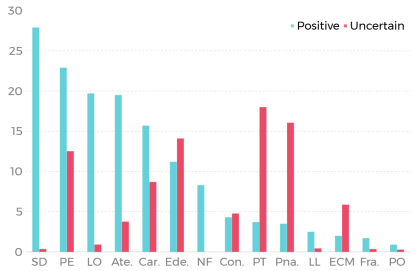


Figure 4: **Non-stationarity of MIMIC-CXR with its percentage of diseases.** Blue signifies patients with clinically confirmed diagnoses showing the long-tailed characteristic, while red demarcates suspected cases emphasizing the inherent uncertainty within medicine.

We selected the MIMIC-CXR [11] dataset not only for its well-established benchmark enabling rigorous performance evaluation on real-world downstream medical tasks, but also due to its authentic clinical representation that exhibits inherent non-stationarity, particularly long-tail and diagnostic ambiguity. As illustrated in Fig.4, the statistical profiling of 14 thoracic pathologies in MIMIC-CXR reveals dual clinical characteristics: the number of confirmed diseases showed a clear long-tail distribution, and each disease had a large number of uncertain patients. Beyond that, we found that 40.87% of the patients suffered from two or more diseases, with nearly 19.97% experiencing three or more. They all reflect the non-stationary environments of our experimental setup within MIMIC-CXR as the training dataset for reinforced custom-tuning.

In terms of the model, we employ Qwen2.5-VL (7B) [16] to perform supervised fine-tuning (SFT) and reinforced fine-tuning (RFT), cascadedly. And they only train one epoch with a batch size of 2.

More detailed experimental implementations are given in Appendix C.

3.1 Taming the Non-stationary Custom-Tuning

First, to explicitly demonstrate the superior performance of our proposed method in non-stationary environments, especially in robustness, we compare it with other models on MS-CXR-T [17], where instances are chosen from the public MIMIC-CXR. As exhibited in Table 2, our counterfactual preference optimization approach achieve the superior overall performance of 81.8%, surpassing the second CoCa-CXR [18] nearly 12.6%. It demonstrates the robustness of our approach to reinforced fine-tuning in non-stationary environments. While our approach trails TempA-VLP [19] by 1.8% on edema (Ede.) detection, we argue that this performance gap emerges from the utilization of additional annotations from Chest ImaGenome [20] in addition to standard MIMIC-CXR. In terms of the pneumothorax (Pnx.), SFT has achieved a high result of 95.9%, so the slight decrease in CPO does not affect the overall performance of the model.

Notably, supervised fine-tuning (SFT) exhibits suboptimal performance on the clinically correlated diseases, namely consolidation (Con.) and pneumonia (Pna.), as presented in Table 1 of Section.2.3. It empirically validates our Observation 1.1 that the inherent concept drift in disease attribute representation within chain-of-thought reasoning introduces systematic prediction biases.

Beyond that, the substantial performance gains of 22.8% and 17.4% in CPO for consolidation (Con.) and pneumonia (Pna.), respectively, underscore our core contribution, which disentangles the concept drift of CoT in reinforcement learning under non-stationary custom-tuning, achieving robust reasoning.

In terms of DPO[15], while our proposed CPO and DPO share a preference-style optimization framework, CPO is fundamentally distinct. DPO contrasts human-preferred with dispreferred responses, while CPO contrasts factuals with counterfactuals generated under explicit causal interventions, specifically designed to isolate the causal effect. Crucially, as shown in Table 2, our proposed CPO is significantly superior to DPO with the same number of training data, proving they are not functionally equivalent. Besides, We also tried to use GRPO [25] to train Qwen2.5-vl and DeepSeek-VL2 on MIMIC-CXR, but both encountered reward collapse that led to failed training. Inspired by DAPO paper[26], we attribute this to GRPO's reliance on sparse final-answer rewards, where suboptimal reward assignment obscures high-quality samples. In contrast, CPO provides denser causal trajectories, significantly improving generalization and robustness by operating controlled counterfactual interventions.

	Venue	Con.	PE	Pna.	Pnx.	Ede.	Avg.
CTrans [21]	CVPR'23	44.0	61.3	45.1	31.5	65.5	49.5
CheXRelNet [22]	MICCAI'22	47.0	47.0	47.0	36.0	49.0	45.2
BioViL [23]	ECCV'22	56.0	63.0	60.2	42.5	67.5	57.8
BioViL-T [21]	CVPR'23	61.1	67.0	61.9	42.6	68.5	60.2
Med-ST [24]	ICML'24	60.6	67.4	58.5	65.0	54.2	61.1
TempA-VLP [19]	WACV'25	65.2	59.4	73.4	43.1	77.1	63.6
CoCa-CXR [18]	Arxiv'25	70.4	69.6	61.4	72.8	71.8	69.2
SFT		54.9	71.7	70.0	95.9	76.5	73.8
DPO	This paper	63.2	72.4	76.7	93.5	76.3	76.4
CPO		**77.7**	**72.7**	**87.4**	**95.8**	**75.3**	**81.8**

Table 2: **Evaluation results of multi-label chest diseases classification on MS-CXR-T.** Top-1 accuracy is applied to evaluate the performance of different methods. The best-performing models are highlighted in red, with the second-best in blue. SFT denotes the results of supervised fine-tuning, and DPO indicates the direct preference optimization with random negative samples, while the CPO represents our counterfactual preference optimization method.

3.2 Concept Drift-Aware CoT for Accurate Reasoning

	Venue	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	CIDEr
R2Gen [27]	EMNLP'20	0.353	0.218	0.145	0.103	0.277	0.142	-
PPKED [28]	CVPR'21	0.360	0.224	0.149	0.106	0.284	0.149	0.237
AlignTrans [29]	MICCAI'2	0.378	0.235	0.156	0.112	0.283	0.158	-
CMCL [30]	ACL'21	0.344	0.217	0.14	0.097	0.281	0.133	-
Clinical-BERT [31]	AAAI'22	0.383	0.230	0.151	0.106	0.275	0.144	0.151
METransformer [32]	CVPR'23	0.386	0.250	0.169	0.124	0.291	0.152	0.362
DCL [33]	CVPR'23	-	-	-	0.109	0.284	0.150	0.281
R2GenGPT [34]	MetaRad'23	0.408	0.256	0.174	0.125	0.285	0.167	0.244
PromptMRG [35]	AAAI'24	0.398	-	-	0.112	0.268	0.157	-
BtspLLM [36]	AAAI'24	0.402	0.262	0.18	0.128	0.291	0.175	-
MambaXray [37]	Arxiv'24	0.422	0.268	0.184	0.133	0.289	0.167	0.241
CPO	This paper	**0.426**	**0.288**	**0.186**	**0.155**	**0.321**	**0.236**	**0.375**

Table 3: Evaluation results of diagnostic report generation on MIMIC-CXR with various metrics including BLEU-1/-2/-3/-4, ROUGE-L, METEOR and CIDEr. The best-performing models are highlighted in red, with the second-best in blue.

Beyond the classification, we verify our main contribution of accurate reasoning, preserving the beneficial CoT within domain adaptation, while eliminating harmful concept drift. As exhibited in the Table 3, the experiments of diagnostic report generation on MIMIC-CXR are conducted to assess the performance of the thinking in our proposed model with chain-of-thought. Our evaluation

combines multiple metrics: BLEU evaluates terminology accuracy with higher-order scores indicating logical coherence in clinical reasoning, ROUGE-L assesses completeness through narrative alignment, METEOR enables synonym-aware lexical matching, and CIDEr prioritizes clinical details via corpus-informed weighting.

The experimental findings demonstrate that our reasoning framework achieves prominent performance across all evaluation metrics, with particularly notable improvements in BLEU-4 (16.5% improvement), ROUGE-L (10.3% increase) and METEOR (34.8% enhancement) scores, indicating the coherence, completeness, and professionalism of our model's thinking. We attribute it to the enhanced fidelity of our reasoning chains in combating concept drift during non-stationary reinforcement learning processes. These enhancements reveal that our method's superior accuracy stems from its capacity to maintain coherent reasoning pathways even when faced with dynamically shifting environmental parameters of complex RL scenarios.

3.3 Generalized Reinforced Custom-tuning

Method	Open-I	PadChest	PadChest20	ChestXray14	ChestXpert	ChestXDet10
MedCLIP [38]	55.1	50.8	50.1	56.4	74.4	57.1
BiomedCLIP [39]	57.7	51.3	51.0	63.9	67.7	63.0
GLoRIA [40]	58.9	56.5	55.8	61.0	75.0	64.5
BioViL [21]	70.2	65.5	60.8	72.9	78.9	70.8
CheXzero [41]	75.9	62.9	68.8	72.6	87.9	71.3
MedKLIP [42]	75.9	62.9	68.8	72.6	87.9	71.3
KAD [43]	80.7	75.0	73.5	78.9	90.5	73.5
CARZero [44]	83.8	81.0	83.7	81.1	92.3	79.6
CPO	**84.4**	**82.0**	**85.1**	**81.7**	**92.5**	**80.1**

Table 4: **Evaluation results of zero-shot diseases classification on Open-I[45], PadChest[46], PadChest20 [46], ChestXray14 [47], ChestXpert [48] and ChestXDet10 [49].** AUC is applied to evaluate the performance of different methods. The best-performing models are highlighted in red, with the second-best in blue.

Furthermore, we validated the generalization of our model on downstream tasks with zero-shot multi-label classification across six different benchmarks, as presented in Table 4. Experimental results demonstrate that our CPO-driven MLLMs achieve zero-shot superiority over the second-best baseline CARZero [44] across all benchmark datasets, underscoring our remarkable robustness and generalization capabilities even when trained under non-stationary environmental regimes.

3.4 Ablation Study on Inherent Compatibility of CPO and CoT: Two Peas in a Pod

Moreover, we conduct ablation experiments on MIMIC-CXR to validate the feasibility and coordination of the chain-of-thought (CPO) and counterfactual preference optimization (CPO) within reinforced fine-tuning (RFT) in non-stationary environments, as presented in Table 5. Among them, only CoT without CPO in RFT represents the utilization of direct preference optimization (DPO) [15] to drive MLLMs for reinforcement learning. While, the only CPO in RFT denotes the reinforced fine-tuning only applies the diagnosis results without the thinking process during the training.

SFT	RFT		Con.	PE	Pna.	Pnx.	Ede.	Avg.
	CoT	CPO						
✓	-	-	54.9	71.7	70.0	95.9	76.5	73.8
✓	✓	-	58.4	71.2	75.0	94.4	75.5	74.9
✓	-	✓	70.5	72.7	77.3	95.2	75.8	78.3
✓	✓	✓	77.7	72.7	87.4	95.8	75.3	81.8

Table 5: **Ablation evaluation results on chain-of-thought (CoT) and counterfactual preference (CPO) within reinforced fine-tuning (RFT) on MIMIC-CXR**, where all RFT stages follow the supervised fine-tuning (SFT). The ✓denotes that the results are trained with the corresponding module. The results are based on the test split of the MS-CXR-T, with Top-1 accuracy (Acc) as the metric.

The experimental analysis reveals CPO's superior performance gain in reinforcement learning (4.5% vs. CoT's 1.1%). We argue

that it is mainly attributable to its mechanism of introducing causally attributed negative samples that enable decision boundary refinement in feature space, whereas CoT primarily operates through stepwise cognitive scaffolding via enhanced positive samples.Therefore, the inherent synergistic compatibility between CoT and CPO emerges through their complementary roles in reinforcement learning frameworks, with CoT generating reinforcement-aligned positive exemplars and CPO providing causality-attuned negative specimens, jointly orchestrating MLLM training optimization as empirically validated through comprehensive benchmarking results achieving state-of-the-art performance.

4 Conclusion and Outlooks

In this paper, we present counterfactual preference optimization (CPO), a novel, robust and generalized reinforced custom-tuning paradigm tailored for non-stationary environments. We employ concept drift theory to methodically formalize the bias within the autoregressive token generation of MLLMs and put forward a causal counterfactual thinking to mitigate these detrimental drifts and keep good domain adaptation. By virtue of this framework, CPO is devised to counteract the unpredictable distribution changes occurring within non-stationary environments.

We hope that our work will inspire future advancements in counterfactual cause of reinforced learning paradigm, specifically addressing biases originating from real-world data challenges. In future research, we will focus on the efficiency of counterfactual causes in reinforced fine-tuning, and broader applications.

Acknowledgment

The work was supported by the Australian Research Council (ARC) under Laureate project FL190100149.

References

[1] Trung, L., X. Zhang, Z. Jie, et al. ReFT: Reasoning with Reinforced Fine-Tuning. In L.-W. Ku, A. Martins, V. Srikumar, eds., Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7601–7614. Association for Computational Linguistics, Bangkok, Thailand, 2024.

[2] Liu, Z., Z. Sun, Y. Zang, et al. Visual-RFT: Visual Reinforcement Fine-Tuning, 2025.

[3] Alayrac, J.-B., J. Donahue, P. Luc, et al. Flamingo: A Visual Language Model for Few-Shot Learning. Advances in Neural Information Processing Systems, 35:23716–23736, 2022.

[4] Bai, J., S. Bai, S. Yang, et al. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023.

[5] Dai, W., J. Li, D. Li, et al. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning, 2023.

[6] Lu, J., A. Liu, F. Dong, et al. Learning under Concept Drift: A Review. 31(12):2346–2363, 2019.

[7] Yang, X., J. Lu, E. Yu. Adapting multi-modal large language model to concept drift from pre-training onwards. arXiv preprint arXiv:2405.13459, 2025.

[8] Pearl, J. Causal diagrams for empirical research. Biometrika, 82(4):669–688, 1995.

[9] —. Causal inference in statistics: a primer. John Wiley & Sons, 2016.

[10] Yang, X., J. Lu, E. Yu. Causal-informed contrastive learning: Towards bias-resilient pre-training under concept drift. arXiv preprint arXiv:2502.07620, 2025.

[11] Johnson, A. E., T. J. Pollard, S. J. Berkowitz, et al. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data, 6(1):317, 2019.

[12] Pearl, J. Direct and indirect effects. In Probabilistic and causal inference: the works of Judea Pearl, pages 373–392. 2022.

[13] Zhang, X., C. Wu, Y. Zhang, et al. Knowledge-enhanced visual-language pre-training on chest radiology images. 14(1):4542, 2023.

[14] Zhou, X., X. Zhang, C. Wu, et al. Knowledge-enhanced Visual-Language Pretraining for Computational Pathology, 2024.

[15] Rafailov, R., A. Sharma, E. Mitchell, et al. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. 36:53728–53741, 2023.

[16] Team, Q. Qwen2.5-vl, 2025.

[17] Bannur, S., S. Hyland, F. Liu, et al. Learning to exploit temporal structure for biomedical vision-language processing. In The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023.

[18] Chen, Y., S. Xu, A. Sellergren, et al. Coca-cxr: Contrastive captioners learn strong temporal structures for chest x-ray vision-language understanding, 2025.

[19] Yang, Z., L. Shen. Tempa-vlp: Temporal-aware vision-language pretraining for longitudinal exploration in chest x-ray image. In 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 4625–4634. 2025.

[20] Wu, J. T., N. N. Agu, I. Lourentzou, et al. Chest imagenome dataset for clinical reasoning. arXiv preprint arXiv:2108.00316, 2021.

[21] Bannur, S., S. Hyland, Q. Liu, et al. Learning to exploit temporal structure for biomedical vision-language processing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 15016–15027. 2023.

[22] Karwande, G., A. B. Mbakwe, J. T. Wu, et al. Chexrelnet: An anatomy-aware model for tracking longitudinal relationships between chest x-rays. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 581–591. Springer, 2022.

[23] Boecking, B., N. Usuyama, S. Bannur, et al. Making the most of text semantics to improve biomedical vision–language processing. In European conference on computer vision, pages 1–21. Springer, 2022.

[24] Yang, J., B. Su, X. Zhao, et al. Unlocking the power of spatial and temporal information in medical multimodal pre-training. In Forty-first International Conference on Machine Learning. 2024.

[25] Shao, Z., P. Wang, Q. Zhu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300, 2024.

[26] Yu, Q., Z. Zhang, R. Zhu, et al. Dapo: An open-source llm reinforcement learning system at scale. arXiv preprint arXiv:2503.14476, 2025.

[27] Chen, Z., Y. Song, T.-H. Chang, et al. Generating radiology reports via memory-driven transformer. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1439–1449. 2020.

[28] Liu, F., X. Wu, S. Ge, et al. Exploring and distilling posterior and prior knowledge for radiology report generation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 13753–13762. 2021.

[29] You, D., F. Liu, S. Ge, et al. Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24, pages 72–82. Springer, 2021.

[30] Liu, F., S. Ge, X. Wu. Competence-based multimodal curriculum learning for medical report generation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3001–3012. 2021.

[31] Yan, B., M. Pei. Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation. In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pages 2982–2990. 2022.

[32] Wang, Z., L. Liu, L. Wang, et al. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11558–11567. 2023.

[33] Li, M., B. Lin, Z. Chen, et al. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3334–3343. 2023.

[34] Wang, Z., L. Liu, L. Wang, et al. R2gengpt: Radiology report generation with frozen llms. Meta-Radiology, 1(3):100033, 2023.

[35] Jin, H., H. Che, Y. Lin, et al. Promptmrg: Diagnosis-driven prompts for medical report generation. In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pages 2607–2615. 2024.

[36] Liu, C., Y. Tian, W. Chen, et al. Bootstrapping large language models for radiology report generation. In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pages 18635–18643. 2024.

[37] Wang, X., F. Wang, Y. Li, et al. CXPMRG-Bench: Pre-training and Benchmarking for X-ray Medical Report Generation on CheXpert Plus Dataset, 2024.

[38] Wang, Z., Z. Wu, D. Agarwal, et al. Medclip: Contrastive learning from unpaired medical images and text. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, vol. 2022, page 3876. 2022.

[39] Zhang, S., Y. Xu, N. Usuyama, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915, 2023.

[40] Huang, S.-C., L. Shen, M. P. Lungren, et al. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In Proceedings of the IEEE/CVF international conference on computer vision, pages 3942–3951. 2021.

[41] Tiu, E., E. Talius, P. Patel, et al. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. Nature biomedical engineering, 6(12):1399–1406, 2022.

[42] Wu, C., X. Zhang, Y. Zhang, et al. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 21372–21383. 2023.

[43] Zhang, X., C. Wu, Y. Zhang, et al. Knowledge-enhanced visual-language pre-training on chest radiology images. Nature Communications, 14(1):4542, 2023.

[44] Lai, H., Q. Yao, Z. Jiang, et al. Carzero: Cross-attention alignment for radiology zero-shot classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11137–11146. 2024.

[45] Demner-Fushman, D., S. Antani, M. Simpson, et al. Design and development of a multimodal biomedical information retrieval system. Journal of Computing Science and Engineering, 6(2):168–177, 2012.

[46] Bustos, A., A. Pertusa, J.-M. Salinas, et al. Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical image analysis, 66:101797, 2020.

[47] Wang, X., Y. Peng, L. Lu, et al. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017.

[48] Irvin, J., P. Rajpurkar, M. Ko, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI conference on artificial intelligence, vol. 33, pages 590–597. 2019.

[49] Liu, J., J. Lian, Y. Yu. Chestx-det10: Chest x-ray dataset on detection of thoracic abnormalities, 2020.

[50] Lu, J., A. Liu, Y. Song, et al. Data-driven decision support under concept drift in streamed big data. Complex & intelligent systems, 6(1):157–163, 2020.

[51] Wang, K., L. Xiong, A. Liu, et al. A self-adaptive ensemble for user interest drift learning. 577:127308, 2024.

[52] Jiao, B., Y. Guo, D. Gong, et al. Dynamic Ensemble Selection for Imbalanced Data Streams With Concept Drift. 35(1):1278–1291, 2024.

[53] Cerqueira, V., H. M. Gomes, A. Bifet, et al. STUDD: A student–teacher method for unsupervised concept drift detection. 112(11):4351–4378, 2023.

[54] Yang, X., Y. Chen, X. Yue, et al. T-distributed Spherical Feature Representation for Imbalanced Classification. Proceedings of the AAAI Conference on Artificial Intelligence, 37(9):10825–10833, 2023.

[55] Yu, E., J. Lu, B. Zhang, et al. Online boosting adaptive learning under concept drift for multistream classification. In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pages 16522–16530. 2024.

[56] Yu, E., Y. Song, G. Zhang, et al. Learn-to-adapt: Concept drift adaptation for hybrid multiple streams. 496:121–130, 2022.

[57] Yu, H., W. Liu, J. Lu, et al. Detecting group concept drift from multiple data streams. 134:109113, 2023.

[58] Liu, W., X. Yue, Y. Chen, et al. Trusted multi-view deep learning with opinion aggregation. In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pages 7585–7593. 2022.

[59] Liu, W., Y. Chen, X. Yue. Enhancing testing-time robustness for trusted multi-view classification in the wild. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 15508–15517. 2025.

[60] —. Building trust in decision with conformalized multi-view deep classification. In Proceedings of the 32nd ACM International Conference on Multimedia, pages 7278–7287. 2024.

[61] Li, W., X. Yang, W. Liu, et al. DDG-DA: Data Distribution Generation for Predictable Concept Drift Adaptation. 36(4):4092–4100, 2022-06-28.

[62] Yang, X., H. Zhang, J. Cai. Deconfounded Image Captioning: A Causal Retrospect. 45(11):12996–13010, 2023.

[63] Liu, B., D. Wang, X. Yang, et al. Show, Deconfound and Tell: Image Captioning With Causal Inference. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18041–18050. 2022.

[64] Zhang, C., L. Zhang, D. Zhou. Causal Walk: Debiasing Multi-Hop Fact Verification with Front-Door Adjustment, 2024.

[65] Choi, S., M. Jeong, H. Han, et al. C2L: Causally Contrastive Learning for Robust Text Classification. 36(10):10526–10534, 2022.

[66] Rohekar, R. Y., Y. Gurwicz, S. Nisimov. Causal Interpretation of Self-Attention in Pre-Trained Transformers. 36:31450–31465, 2023.

[67] Yang, X., H. Zhang, G. Qi, et al. Causal Attention for Vision-Language Tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9847–9857. 2021.

[68] Yang, X., L. Xu, H. Li, et al. One leaf reveals the season: Occlusion-based contrastive learning with semantic-aware views for efficient visual representation. In Forty-second International Conference on Machine Learning. 2025.

[69] Christiano, P. F., J. Leike, T. Brown, et al. Deep reinforcement learning from human preferences. Advances in neural information processing systems, 30, 2017.

[70] Ouyang, L., J. Wu, X. Jiang, et al. Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35:27730–27744, 2022.

[71] Jaech, A., A. Kalai, A. Lerer, et al. Openai o1 system card. arXiv preprint arXiv:2412.16720, 2024.

[72] Bai, Y., S. Kadavath, S. Kundu, et al. Constitutional ai: Harmlessness from ai feedback. arXiv preprint arXiv:2212.08073, 2022.

[73] Guo, D., D. Yang, H. Zhang, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948, 2025.

[74] Gulcehre, C., T. L. Paine, S. Srinivasan, et al. Reinforced self-training (rest) for language modeling, 2023.

[75] Yuan, Z., H. Yuan, C. Li, et al. Scaling relationship on learning mathematical reasoning with large language models, 2024.

[76] Zeng, W., Y. Huang, L. Zhao, et al. B-STar: Monitoring and balancing exploration and exploitation in self-taught reasoners. In The Thirteenth International Conference on Learning Representations. 2025.

[77] Zhang, Z., C. Zheng, Y. Wu, et al. The lessons of developing process reward models in mathematical reasoning. arXiv preprint arXiv:2501.07301, 2025.

[78] Yang, X., L. Xu, H. Sun, et al. Enhancing visual grounding and generalization: A multi-task cycle training approach for vision-language models. arXiv preprint arXiv:2311.12327, 2024.

[79] Yang, X., J. Lu, E. Yu. Learning from all: Concept alignment for autonomous distillation from multiple drifting mllms. arXiv preprint arXiv:2510.04142, 2025.

[80] Li, Z.-Z., D. Zhang, M.-L. Zhang, et al. From system 1 to system 2: A survey of reasoning large language models. arXiv preprint arXiv:2502.17419, 2025.

[81] Chen, Q., L. Qin, J. Liu, et al. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. arXiv preprint arXiv:2503.09567, 2025.

[82] Zhang, Z., A. Zhang, M. Li, et al. Multimodal chain-of-thought reasoning in language models. arXiv preprint arXiv:2302.00923, 2023.

[83] Chen, Q., L. Qin, J. Zhang, et al. m^3 cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. arXiv preprint arXiv:2405.16473, 2024.

[84] Wang, Y., S. Wu, Y. Zhang, et al. Multimodal chain-of-thought reasoning: A comprehensive survey. arXiv preprint arXiv:2503.12605, 2025.

[85] Zheng, G., B. Yang, J. Tang, et al. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. Advances in Neural Information Processing Systems, 36:5168–5191, 2023.

[86] Liu, W., Y. Chen, X. Yue, et al. Enhancing reliability in medical image classification of imperfect views. IEEE Transactions on Circuits and Systems for Video Technology, 2025.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

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

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

Justification: We discuss limitations of our approach and the outlook of further work in the section of conclusions.

Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: We provide the full set of assumptions and a complete proof in the main manuscript and the supplemental.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We provide full information about the replication experiments in the main manuscript and make our code and data publicly available.

Guidelines:

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

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

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

   Answer: [Yes]

   Justification: We made our code and data publicly available on github as anonymous. The anonymous link is `https://anonymous.4open.science/r/CPO-FD61/`.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: We present the details of the experiments in the supplemental material.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: We report results of experiments with statistical significance.

   Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We provide the information about computer resources in the supplemental material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: We made our code and data publicly as anonymous in the review stage.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We discuss the broader impacts of our work in the section of conclusions.

    Guidelines:

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

11. **Safeguards**

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

    Answer: [NA]

    Justification: We have no such risks.

Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: We cite the original paper that produced the code package or dataset.

    Guidelines:

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

13. **New assets**

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

    Answer: [Yes]

    Justification: We have talked about the details of the dataset and code as part of our submissions.

    Guidelines:

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

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

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

Answer: [NA]

Justification: We do not involve crowdsourcing nor research with human subjects.

Guidelines:

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

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

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

Answer: [NA]

Justification: We do not involve crowdsourcing nor research with human subjects.

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Appendix

A Related Works

A.1 Concept Drift

In their survey spanning multiple studies, Lu et al. [6, 50] establish a comprehensive taxonomy of concept drift mitigation strategies, categorizing prevailing approaches into three principal paradigms: error rate-driven adaptations [51, 52], data distribution-aware methodologies [7, 53, 54], and multi-hypothesis frameworks [55, 56]. Our work aligns with the distribution-centric paradigm, which distinguishes itself through its dual capacity for both precise drift detection via rigorous statistical analysis and holistic drift characterization across temporal, spatial, and quantitative dimensions. Specifically, these distribution-driven techniques enable not merely the identification of concept drift occurrence but also facilitate granular diagnostics through temporal localization of drift emergence, feature subspace attribution, and severity quantification - capabilities that render them particularly advantageous for developing interpretable adaptive systems requiring both drift awareness and targeted model recalibration.

Recent advances in concept drift adaptation have yielded sophisticated methodologies across diverse learning scenarios. The Online Boosting Adaptive Learning (OBAL) framework [55] has emerged as a dual-phase solution for multistream classification challenges, initially employing Adaptive Covariate Shift Adaptation (AdaCOSA) to model dynamic inter-stream correlations before transitioning to Gaussian Mixture Model-based weighting for asynchronous drift mitigation. Complementing this, CDMLLM [7] reveals critical vulnerabilities in vision-language models through systematic analysis of concept drift-induced biases across pre-training and fine-tuning stages, proposing a unified framework that synergizes T-distribution adaptation for long-tailed calibration with explicit out-of-distribution detection to enhance multimodal alignment robustness. Expanding the scope beyond individual data streams, GDDM [57] introduces a distribution-free statistical framework for detecting subtle group-level concept drifts in multi-stream environments through adaptive hypothesis testing mechanisms. Additionally, the studies by [58–60] introduce a multi-perspective uncertainty framework designed to handle concept drift in diverse data streams. This approach utilizes set-based prediction methods to unify probabilistic results into clear categorical formats. In parallel, DDG-DA [61] pioneers anticipatory concept drift adaptation by modeling environmental evolution through predictive factor analysis and synthetic data generation, effectively bridging current observations with projected distribution shifts. Advancing unsupervised detection paradigms, STUDD [53] establishes a teacher-student discrepancy framework that leverages predictive consistency analysis to enable label-agnostic drift identification while maintaining detection sensitivity, thereby addressing practical deployment constraints in evolving data environments.

A.2 Causal Inference

Recently, increasing researchers have incorporated causal inference into deep-learning models, especially in large models. Deconfounded Image Captioning (DIC) [62] is proposed to address dataset bias in vision-language models through a causal lens, that integrates backdoor and front-door adjustments for systematic bias mitigation. The framework provides principled causal analysis of spurious correlations in multimodal alignment, offering theoretical grounding for decomposing bias sources through structured interventions. Likewise, aiming for spurious correlations induced by visual and linguistic biases during training, CIIC [63] is proposed as a causal intervention framework combining an Interventional Object Detector (IOD) and Interventional Transformer Decoder (ITD) guided by structural causal models. By applying backdoor adjustment through IOD's feature disentanglement and ITD's dual de-confounding mechanism, their approach systematically mitigates confounding effects across encoding and decoding stages, demonstrating enhanced generalization through causal correlation modeling. Similarly, targeting multi-hop fact verification bias in the large language model, Causal Walk [64] is proposed, a front-door adjustment framework that disentangles complex spurious correlations in evidence chains. The method models reasoning paths as mediators in structural causal models, decomposing causal effects via random walk-based treatment-mediator estimation and geometric mean-based mediator-outcome approximation. By integrating adversarial and symmetric datasets synthesized with large language models, the approach demonstrates superior debiasing performance.

Recent advances in causal representation learning have produced innovative methodologies to address confounding biases in large models. The C2L framework [65] tackles model fragility through contrastive counterfactual synthesis, introducing a collective decision mechanism that aggregates predictions across probabilistically generated counterfactual sets while enforcing causal invariance via distributional consensus supervision, thereby overcoming dataset-inherent bias limitations of conventional augmentation approaches. Building on causal interpretability, ABCD [66] establishes formal theoretical grounding for Transformer architectures by reinterpreting self-attention mechanisms as structural equation estimators that capture conditional independence relations through partial correlation analysis in deep attention layers, enabling zero-shot causal discovery over input sequences while accounting for latent confounders through repurposed pre-trained models. Expanding the causal intervention paradigm, Causal Attention (CATT) [67] implements front-door adjustment via dual-path processing of In-Sample and Cross-Sample Attention, strategically integrating external contextual information through CS-ATT while preserving standard attention mechanisms to dynamically mitigate spurious correlations without explicit confounder specification, thereby achieving bias-resistant vision-language alignment through implicit causal disentanglement. Moreover, ResilientCL [10, 68] proposes the causal interventional contrastive objective to mitigate the concept drift within the momentum network of contrastive pre-training paradigm.

A.3 Reinforced Fine-tuning

The integration of reinforcement learning (RL) into post-training alignment of large language models (LLMs) has undergone remarkable evolution since OpenAI's seminal work on Reinforcement Learning from Human Feedback (RLHF) [69], which established a foundational paradigm for aligning model outputs with human values [70]. While early implementations like OpenAI-o1 [71] demonstrated the efficacy of human preference modeling, the prohibitive costs of manual annotation have catalyzed a paradigm shift toward automated reward generation through pre-trained systems. This transition has yielded innovative methodologies ranging from Bai et al.'s [72] constitutional approach utilizing sparse natural language feedback as proxy signals, to DeepSeek's progressive framework that first established baseline performance through pure RL (R0) before introducing their R1 variant [73]. The latter achieved enhanced generalization through cyclic alternation between supervised fine-tuning and their novel GRPO optimization protocol [25], exemplifying the field's progression toward self-contained alignment systems.

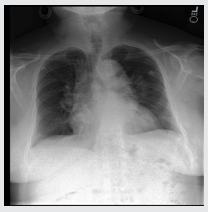
Besides, the landscape of alignment methodologies continues to diversify through innovative paradigms: ReST [74] employs iterative self-generation of policy-derived samples to refine LLMs via offline reinforcement learning, while DPO [15] fundamentally reformulates alignment as direct preference optimization through implicit reward modeling. Concurrent developments span Rejection Sampling Fine-Tuning's [75] curation of validated reasoning trajectories for supervised augmentation, and ReFT's [1] phased optimization combining SFT initialization with PPO-driven exploration of automated reasoning path generation. Building upon these foundations, Visual-RFT [2] extends GRPO-based strategies to multimodal contexts, enhancing visual-language alignment under data scarcity, whereas B-STaR [76] introduces dynamic configuration adaptation for self-teaching systems through principled exploration-exploitation balancing. Pushing the boundaries of evaluation rigor, Qwen-Math-PRM [77] synergizes Monte Carlo estimation with LLM-as-judge consensus filtering while pioneering a hierarchical assessment framework integrating stepwise and holistic performance metrics. Moreover, ViLaM [78] performs visual grounding unsupervised via reinforced learning under the open-world environment. Besides, APO [79] leverages reinforcement learning to align the knowledge in multiple teacher models for distillation.

A.4 Multimodal Reasoning

Recent advances in Long Chain-of-Thought (Long CoT) reasoning [80, 81] have significantly enhanced the capacity of Large Language Models (LLMs) to perform multi-step reasoning and self-correction. By incorporating self-reflection strategies, these models can dynamically diagnose and revise their intermediate reasoning traces, thereby mitigating certain types of reasoning inconsistencies during inference. In contrast, our approach introduces a proactive intervention at the training stage through counterfactual sample generation, which explicitly shapes the model's causal representations and decision boundaries. This enables the model to internalize more consistent causal reasoning patterns from the outset, rather than relying on post-hoc correction during inference. Hence, we view our method and Long CoT-based self-reflection strategies as complementary: while Long CoT

enhances reasoning reliability at runtime, our approach strengthens causal robustness during learning. We believe future work could explore integrating both directions—leveraging Long CoT's reflective inference mechanisms together with counterfactual training—to further advance consistent causal reasoning in reasoning-oriented LLMs.

In parallel, Multimodal Chain-of-Thought (MM-CoT) reasoning [82–85] has emerged as a crucial paradigm for enabling LLMs to align and reason across heterogeneous modalities such as vision and language. For instance, M3CoT [83] and Ddcot [85] introduce multi-domain, multi-step reasoning frameworks that emphasize structured cross-modal inference, while the survey in [84] provides a comprehensive taxonomy of MM-CoT paradigms and benchmarks. Compared with these inference-time reasoning approaches, our method introduces a proactive causal intervention during training through counterfactual sample generation, aiming to enhance causal robustness within multimodal reasoning processes. We view our approach as complementary to both Long CoT and MM-CoT paradigms—while they improve the reasoning trajectory at inference time, ours focuses on stabilizing causal representations during model optimization. This synergy could be a promising direction for future research.



Ground Truth:

Findings:
PA and lateral views of the chest demonstrate low lung volumes. Tiny bilateral pleural effusions are new since ___. No signs of pneumonia or pulmonary vascular congestion. Heart is top normal in size though this is stable. Aorta is markedly tortuous, unchanged. Aortic arch calcifications are seen. There is no pneumothorax. No focal consolidation. Partially imaged upper abdomen is unremarkable.

Diagnosis:
The disease of this patient is Pleural Effusion.

Counterfactual Thinking:

Findings:
PA and lateral views of the chest demonstrate low lung volumes with new signs of consolidation in both lower lobes. Tiny bilateral pleural effusions are noted. No signs of pneumonia were previously present. There is evidence of pulmonary vascular congestion. The heart is slightly enlarged, though this has not been stable. The aorta is markedly tortuous, unchanged. Aortic arch calcifications are seen. There is no pneumothorax. The partially imaged upper abdomen is unremarkable.

Diagnosis:
The disease of this patient is Pneumonia.

Findings:
PA and lateral views of the chest demonstrate low lung volumes. Tiny bilateral pleural effusions are new since last examination. No signs of pneumonia or pulmonary vascular congestion. Heart is top normal in size though this is stable. Aorta is markedly tortuous, unchanged. Aortic arch calcifications are seen. There is no pneumothorax. Focal consolidation is noted in the right lower lobe. Partially imaged upper abdomen is unremarkable.

Diagnosis:
The disease of this patient is Consolidation.

Ground Truth:

Findings:
The patient's overlying arm on the lateral view partially obscures the view and makes evaluation of the lateral view suboptimal. Left greater than right biapical scarring is noted. No pleural effusion or pneumothorax is seen. The cardiac and mediastinal silhouettes are unremarkable. Multiple surgical clips are noted overlying the left hemi thorax and the left axilla. Difficult to assess for medial left clavicular injury, nondisplaced fracture not excluded.

Diagnosis:
The disease of this patient is Lung Opacity.

Counterfactual Thinking:

Findings:
The patient's overlying arm on the lateral view partially obscures the view and makes evaluation of the lateral view suboptimal. Left greater than right biapical scarring is noted. A small left-sided pneumothorax is suspected but difficult to assess definitively due to the arm overlap. No pleural effusion is seen. The cardiac and mediastinal silhouettes are unremarkable. Multiple surgical clips are noted overlying the left hemithorax and the left axilla. Nondisplaced fracture of the medial left clavicle cannot be excluded.

Diagnosis:
The disease of this patient is Pneumothorax.

Findings:
The patient's overlying arm on the lateral view partially obscures the view and makes evaluation of the lateral view suboptimal. Left greater than right biapical scarring is noted with evidence of edema in the left lower lobe. No pleural effusion or pneumothorax is seen. The cardiac and mediastinal silhouettes are unremarkable. Multiple surgical clips are noted overlying the left hemithorax and the left axilla. Difficult to assess for medial left clavicular injury, nondisplaced fracture not excluded.

Diagnosis:
The disease of this patient is Edema.

Figure 5: Samples of CXR-CounterFact (CCF) Dataset.

B CXR-CounterFact (CCF) Dataset

The motivation for choosing chest diagnostics as the application is due to the abundance of public professional medical diagnosis reports [86], which serve as rich CoT reasoning process. This domain provides an ideal platform for studying MLLM reasoning processes and validating the robustness of our proposed CPO in real-world settings.

Figure 5 showcases the samples utilized for training and validation in our study. We use a medical-specific LLM to generate the related caption of the image, with the prompt of:

"This is a radiology chest DR examination report of a patient: <Report>.

This is a diagram of the relationship between lung diseases and their radiographic manifestations: <Concept Graph>

Please generate a counterfactual radiology text showing <disease> based on the relationship and above context, with the same formatting." .

As depicted in Figure 5 , comprehensive descriptions of the image are provided through long-form text, encompassing details such as size, position, relationships, and other relevant information about the disease present in the image. This ensures a detailed and information-rich depiction of the visual content. We have publicly released the datasets used for training and validation.

C Implementation Details

In this section, implementation details are provided.

In terms of the supervised fine-tuning progress, the hyperparameters are presented in Table 6. Qwen2.5-VL (7B) [16] is applied as our pre-trained model. During the SFT, we utilize the AdamW optimizer, which is configured with a cosine annealing schedule as the learning policy. The initial learning rate is set to 1×10^{-4}, and the AdamW optimizer is employed with hyperparameters $\beta = (0.9, 0.98)$. Additionally, we set the weight decay to 0.05 and the dropout rate to 0.1. During the first 20 warm-up steps, the learning rate increases to 1×10^{-4}, and subsequently decays to 10^{-7}. Unless otherwise specified, the supervised fine-tuning of our multi-modal large language model consists of 660 steps, executed on 2×2 NVIDIA A100 GPUs.

Table 6: The training hyperparameters of our MLLM.

Supervised Fine-tuning	
Training Steps	660
Warmup Steps	20
Warmup Ratio	0.05
Optimizer	AdamW
Learning Rate	1e-4
Learning Rate Decay	Cosine
Adam β	(0.9, 0.98)
Weight Decay	0.05
Batch Size	15

Counterfactual Preference Optimzation	
Training Steps	7,750
Warmup Steps	0
Optimizer	AdamW
Learning Rate	2e-5
Learning Rate Decay	Cosine
Adam β	(0.9, 0.98)
Weight Decay	0.05
Batch Size	4

While in the counterfactual preference optimization (CPO), the initial learning rate is reduced to 2×10^{-5} without the warmup. The visual encoder and text decoder are frozen out of the training. Thus, the batch size can be decreased to 4. The reinforced custom-tuning consists of 7,750 steps, executed on 2×2 NVIDIA A100 GPUs. Other training parameters are the same as the fine-tuning.

It is worth noting that, both the SFT and CPO models were trained for exactly one epoch, as suggested by Qwen2.5-vl [4]. The difference in the number of training steps arises solely because the CPO dataset is approximately three times larger than the SFT dataset due to the addition of counterfactual trajectories. Therefore, both models completed their training after seeing their respective datasets once, indicating comparable convergence points in terms of epoch count, rather than SFT being stopped prematurely.

C.1 Computational Cost

The main computational overhead of our framework stems from counterfactual sample generation using LLM experts. To ensure feasibility, we employ a targeted generation strategy based on the hierarchical concept graph, which encodes disease entities, radiographic features, and their statistical dependencies. This allows us to selectively perturb drift-prone features and generate only the two most relevant counterfactual samples per instance, rather than performing random perturbations.

Moreover, all counterfactuals are generated once offline rather than dynamically during training. The construction of the CCF dataset took approximately five days on four A100 GPUs, including both inference and validation. Although this process is computationally intensive, it represents a one-time cost that is affordable for real-world deployment. The subsequent CPO fine-tuning and evaluation were performed on 2×A100 GPUs with negligible extra overhead compared to standard fine-tuning.

C.2 Concept Graph Construction

The prompt we used for automated extraction from the MIMIC-CXR dataset is as follows:

```
TASK ROLE: You are a senior chest radiologist. I will provide a large
    number of chest DR case report texts. Please automatically extract
     key imaging feature words related to various chest diseases from
    these reports, and use these features to build a structured
    imaging knowledge graph to reveal the association between
    different diseases based on common imaging features.

CORE REQUIREMENTS:
1. Standardized feature extraction:
    - Extract all abnormal imaging descriptors from reports.
    - Normalize terminology using Radiology Lexicon (RadLex) and
        Fleischner Society guidelines.
2. Disease-feature mapping:
    - Link standardized features to diagnoses diseases per report.
    - Identical features MUST use identical normalized terms across
        diseases.
3. Knowledge graph construction:
    - Nodes:
        Diseases, such as Pneumothorax, Atelectasis and Pneumonia;
        Features, such as lung opacity and air bronchograms.
    - Relationship:
        (Disease)-[HAS\_FEATURE]->(Feature);
        (Feature)-[ASSOCIATED\_WITH]->(Disease)
    - Semantics:
        Diseases sharing one more identical feature node are
            interconnected.

OUTPUT REQUIREMENTS:
Please output the final constructed radiological knowledge graph and
    the standardized feature-disease association data extracted from
    the report in a structured JSON format. The format is as follows:
{
    "data":[
        {"diseases":[], "features":[]},
        // ... cases
    ],
    "disease_feature_map":{
        "Pneumonia": ["Consolidation", "Ground-glass opacity", "Air
            bronchogram", "Pleural effusion"],
        // ... relationships
    }
}
```