Appendix of Clinically-Grounded Counterfactual Reasoning for Medical Video Diagnosis

A. Additional Details about Colposcopy

A.1. Task Description

Colposcopy is a standard procedure for cervical cancer screening, where the cervix is inspected under sequential reagent applications—saline, acetic acid, alcohol, and iodine—to reveal tissue responses indicative of underlying pathology [6]. During the examination, gynecologists determine biopsy sites, *i.e.*, specific anatomical regions from which tissue samples are taken for histopathological confirmation [9]. Clinically, the cervix is divided into **12 clock-position regions**, and each region may contain one or more suspicious lesions [2]. Expert colposcopists assess reagent-induced color and texture changes such as acetowhitening, mosaicism, punctation, and iodine-negative areas [4]. Among these stages, the transition from acetic-acid reaction to iodine staining provides the most discriminative cues for identifying pathological regions and is therefore the primary basis for biopsy-site localization, as shown in Fig. A1. Accordingly, the task is formulated as a multi-label classification problem that predicts biopsy-site locations from the full multi-stage colposcopy video.

A.2. Training Details

Each examination contains a four-stage colposcopy video. For temporal modeling, the full sequence is divided into overlapping 16-frame clips using a temporal stride of 8. The overlap preserves smooth temporal continuity across stages and ensures that reagent-induced tissue changes are consistently captured for representation learning.

B. Additional Experiment

To assess the generalization ability of **MEDVCR** beyond video-based diagnosis, we evaluate it on a static imaging task, *i.e.*, bilateral mammography classification.

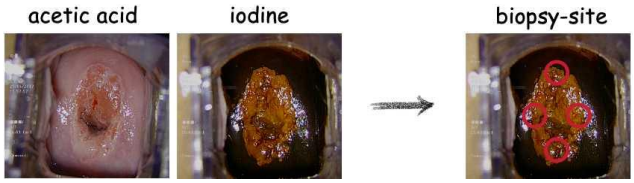

Figure A1. **Illustration of colposcopic biopsy-site determination.** By comparing tissue responses between the acetic-acid and iodine stages, clinicians localize suspicious areas on the cervix (organized into clock-position regions) for targeted biopsy sampling.

Table A1. **Quantitative results of mammography image analysis on INBreast [5].** Results are obtained through five-fold cross-validation. See Sec. B.1 for details.

Category	Methods	AUC ↑
General	ResNet50[CVPR16] [3]	69.0
	CNN-based[MICCAI16] [1]	76.0
Breast-specific	Zhu *et al.*[MICCAI17] [10]	86.0
	Wu *et al.*[TMI19] [8]	86.3
	Wang *et al.*[TIP21] [7]	90.9
MEDVCR (Ours)		**93.4**

B.1. Mammography Image Analysis

Task description Mammography is the primary imaging technique for breast cancer screening and diagnosis [7]. This experiment is designed to verify the generalization capability of the proposed counterfactual reasoning framework. The breasts exhibit natural bilateral symmetry, where lesions on one side rarely appear in the corresponding region of the opposite breast [10]. Therefore, radiologists identify potential malignancies by examining asymmetries between paired views. This task involves detecting pathological lesions based on paired mammograms.

Dataset Experiments are performed on the public INBreast dataset [5], a high-quality benchmark for mammogram analysis. The dataset encompasses diverse lesion types, including masses, calcifications, and architectural distortions, representing challenging clinical scenarios. It contains 410 mammograms from 115 cases, each with paired left–right views and image-level BI-RADS annotations verified by pathological examination. Following standard practice [7], images are labeled as malignant if BI-RADS > 3, and benign otherwise. For bilateral analysis, 91 valid left–right pairs are used for five-fold cross-validation.

Metrics Performance is evaluated using AUC, following [7].

Compared methods Bilateral mammogram analysis is formulated as a binary classification task. Evaluation involves three categories of representative methods.

- *General methods.* General image classification backbones, *i.e.*, ResNet50 [3] and CNN-based method [1], are employed as standard discriminative baselines.

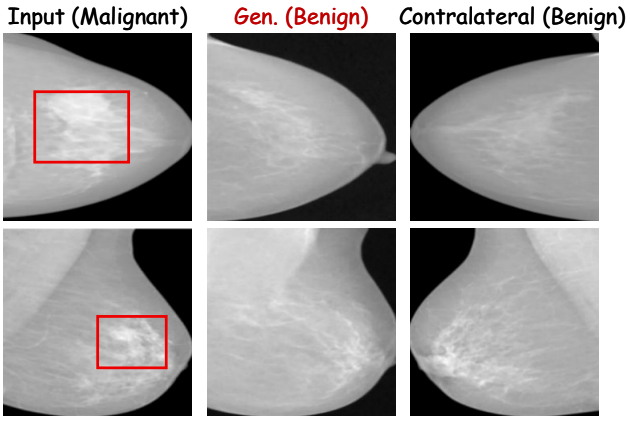

Figure A2. **Visualization of mammography counterfactuals.** Given a malignant input (left), the model generates a benign hypothesis (middle), which closely aligns with the appearance of the true contralateral breast (right). See Sec. B.1 for details.

- *Breast-specific methods.* These methods explicitly model bilateral breast symmetry for classification, including Zhu *et al*. [10], Wu *et al*. [8], and Wang *et al*. [7].
- *Ours.* The proposed framework is adapted for mammography by removing $\mathcal{F}^t$ and retaining only $\mathcal{F}^e$, which independently processes each mammogram image.

Quantitative results Results on INBreast [5] are presented in Tab. A1. **MEDVCR** significantly surpasses the previous best method, Wang *et al*. [7] (*e.g*., 90.9%→**93.4**% in AUC). This demonstrates that our framework generalizes effectively to medical image analysis, validating its versatility across different clinical imaging modalities.

Qualitative results To further examine the behavior of the CG on static imaging, Fig. A2 presents qualitative results on malignant mammograms. For each malignant input (left), the generator produces a benign counterfactual hypothesis (middle). For reference, we also show the true contralateral breast (right), which is typically benign. The generated benign counterfactual suppresses malignant high-density regions and restores coherent glandular and parenchymal patterns. Its morphology resembles the true contralateral breast, demonstrating that our generator preserves global breast architecture while selectively removing pathology-specific structural abnormalities.

C. Discussion

Limitations and future work (i) *Scope of generative modeling.* Our CG is trained only to model short-range transitions between adjacent clinical stages. Its ability to synthesize longer-term or cross-stage evolution has not been evaluated and may limit applicability to procedures with complex temporal dynamics. (ii) *Coverage of clinical knowledge.* The clinical rules incorporated in **MEDVCR** reflect key diagnostic principles but do not encompass the full range of reasoning strategies used by experienced clinicians. Additional domain knowledge or adaptive rule learning may further improve representation fidelity. (iii) *Generality across modalities and institutions.* Although the framework shows strong performance across colposcopy, colonoscopy, and mammography, broader validation on multi-center datasets and diverse imaging modalities is necessary to assess the robustness of **MEDVCR**.

Broader impact This work explores the potential of counterfactual reasoning for medical video diagnosis. By generating clinically plausible benign–malignant hypotheses and explicitly modeling temporal tissue transitions, **MEDVCR** provides more transparent and clinically aligned diagnostic cues than conventional end-to-end methods. We hope this work encourages further development of clinically grounded, interpretable diagnostic systems for medical video analysis.

References

[1] Neeraj Dhungel, Gustavo Carneiro, and Andrew P Bradley. The automated learning of deep features for breast mass classification from mammograms. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2016. A1

[2] Michaela T Hall, Kate T Simms, John M Murray, Adam Keane, Diep TN Nguyen, Michael Caruana, Gigi Lui, Helen Kelly, Linda O Eckert, Nancy Santesso, et al. Benefits and harms of cervical screening, triage and treatment strategies in women living with hiv. *Nature Medicine*, 29(12): 3059–3066, 2023. A1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. A1

[4] Smita Joshi, Richard Muwonge, Ramesh Bhosale, Pritam Chaudhari, Vinay Kulkarni, Mahesh Mandolkar, Kedar Deodhar, Seema Kand, Nikhil Phadke, Shobini Rajan, et al. A randomised controlled non-inferiority trial to compare the efficacy of 'hpv screen, triage and treat'with 'hpv screen and treat'approach for cervical cancer prevention among women living with hiv. *Nature Communications*, 16(1):1888, 2025. A1

[5] Inês C Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria Joao Cardoso, and Jaime S Cardoso. Inbreast: toward a full-field digital mammographic database. *Academic radiology*, 19(2):236–248, 2012. A1, A2

[6] Lena Schreiberhuber, James E Barrett, Jiangrong Wang, Elisa Redl, Chiara Herzog, Charlotte D Vavourakis, Karin Sundström, Joakim Dillner, and Martin Widschwendter. Cervical cancer screening using dna methylation triage in a real-world population. *Nature medicine*, 30(8):2251–2257, 2024. A1

[7] Churan Wang, Jing Li, Fandong Zhang, Xinwei Sun, Hao Dong, Yizhou Yu, and Yizhou Wang. Bilateral asymmetry guided counterfactual generating network for mammogram classification. *Transactions on Image Processing (TIP)*, 30: 7980–7994, 2021. A1, A2

[8] Nan Wu, Jason Phang, Jungkyu Park, Yiqiu Shen, Zhe Huang, Masha Zorin, Stanisław Jastrzebski, Thibault Févry, Joe Katsnelson, Eric Kim, et al. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Transactions on Medical Imaging (TMI)*, 39(4):1184–1194, 2019. A1, A2

[9] Peng Xue, Le Dang, Ling-Hua Kong, Hong-Ping Tang, Hai-Miao Xu, Hai-Yan Weng, Zhe Wang, Rong-Gan Wei, Lian Xu, Hong-Xia Li, et al. Deep learning enabled liquid-based cytology model for cervical precancer and cancer detection. *Nature Communications*, 16(1):3506, 2025. A1

[10] Wentao Zhu, Qi Lou, Yeeleng Scott Vang, and Xiaohui Xie. Deep multi-instance networks with sparse label assignment for whole mammogram classification. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2017. A1, A2