Gastric-X: A Multimodal Multi-Phase Benchmark Dataset for Advancing Vision-Language Models in Gastric Cancer Analysis

(Supplementary File)

A. Abstract

This supplementary document provides extended technical details and additional discussions for the Gastric-X benchmark. Specifically, we present: (1) a description of the multi-phase CT normalization and alignment pipeline in Sec. B; (2) detailed explanations of the provided clinical reports (CT imaging descriptions, endoscopy reports, and diagnostic conclusions) in Sec. C; (3) a description of the creation, verification, and prompting strategy for all VQA pairs in Sec. D; (4) an illustration of the 134 biomedical indicators in Sec. E.

B. Multi-phase CT Standardization Details

Multi-phase CT scans encompass substantial variation across patients and acquisition phases. Our preprocessing pipeline aims to harmonize patterns, standardize geometric properties, and ensure spatial alignment across phases.

Intensity normalization across phases. We apply a unified clipping window of $[-100, 300]$ HU, consistent with recommended gastric soft-tissue imaging ranges. For each CT volume, per-volume z-score normalization is performed after clipping. In addition, we use histogram matching across phases to reduce heterogeneity.

Voxel spacing standardization. All phases are resampled to isotropic spacing of $1.0 \times 1.0 \times 1.0\,\mathrm{mm}^3$ using trilinear interpolation for image data.

Handling different-sized CT slices. Raw scans contain variable numbers of axial slices. Each patient is associated with a coarse 3D bounding region around the stomach, manually annotated by clinical readers. These bounding boxes vary between

$$256 \times 256 \times 160 \text{ and } 288 \times 288 \times 192$$

depending on patient-specific anatomy. Volumes are cropped or padded to the unified shape: $288 \times 288 \times 192$.

Multi-phase alignment. Arterial and delayed phases are rigidly registered to the venous phase using a 6-DOF transformation optimized via mutual information. Registration is implemented with SimpleITK (Elastix backend).

Quality control. Scans with corrupted slices, missing metadata, or excessive misalignment are excluded. Approximately 3–4% of volumes are filtered by this process.

C. The Clinical Reports

Each patient record contains three types of clinical reports:

CT imaging description report. A detailed morphological description authored by radiologists. It covers wall thickening, ulceration, enhancement patterns, perigastric fat infiltration, lymph node size/morphology, and incidental findings.

Endoscopy report. This endoscopy report provides a detailed assessment of the gastrointestinal mucosa, including evaluation of surface texture, ulceration, pit-pattern characteristics, and any other notable structural changes. Lesions are described with explicit documentation of their location, extent, and depth. Biopsy samples, when obtained, are recorded with corresponding anatomical sites to support accurate histologic correlation.

Diagnostic conclusion report. A concise interpretive summary presenting the radiologist's overall impression, including features suggestive of malignancy, estimated TNM staging when applicable, assessment of regional or distant nodal involvement, and any pertinent recommendations for further evaluation or correlation with clinical or pathological findings.

D. The Creation of VQA Pairs

The dataset contains 26,760 VQA pairs derived from clinical reports. Their creation follows a multi-stage pipeline designed to ensure clinical correctness, semantic grounding, and diversity of reasoning patterns.

(1) Large-scale candidate generation using multiple LLMs. Two publicly available large language models, e.g., ChatGPT 4.0, Gemini 2.5 and Claude Sonnet 4.0 were prompted with structured instructions to generate initial question candidates. The prompts targeted clinically meaningful aspects such as lesion characterization, enhancement behavior across phases, staging-relevant findings, anatomical localization, and factual consistency checks. Among all

Table 1. Effectiveness of different prompting strategies for generating clinically valid VQA questions. Validity represents the percentage of Q/A pairs confirmed by both clinicians.

Prompt Type	**Validity (%)**	**Remarks**
Lesion-focused	92.4	Most clinically reliable and consistently grounded.
Staging-focused	88.1	Dependent on level of staging detail documented.
Enhancement-phase	84.7	Sensitive to phase contrast variations.
Localization	79.3	Occasional ambiguity in spatial descriptions.
Yes/No factual	90.5	High factual precision but limited question diversity.

generated candidates, ChatGPT 4.0 contributed 78.32% of the questions that ultimately passed clinical verification.

(2) Prompt design and controlled extraction. To systematically guide generation, we defined five prompt categories: (1) lesion-centric question design, (2) enhancement-phase reasoning, (3) staging-related reasoning, (4) anatomical localization questions, and (5) binary Yes/No factual verification. Each prompt type was designed to reflect reasoning processes typically employed in abdominal radiology. Outputs containing ambiguous phrasing or information not present in the source report were automatically removed.

(3) Final VQA selection and answer fidelity. Only Q/A pairs that strictly adhered to source-report evidence were retained. All answers are derived exclusively from the original CT imaging descriptions or diagnostic conclusions, without augmentation using external medical knowledge.

(4) Double-blind clinical verification. All candidate Q/A pairs underwent sentence-level verification by two independent clinical experts: a radiologist with seven years of experience and a gastroenterology specialist with ten years of experience. Each clinician evaluated the factual correctness of both the question and its corresponding answer by directly comparing them to the source report. Discrepancies were flagged and resolved through consensus. This process ensures that the final VQA set reflects clinically valid reasoning and avoids hallucinated associations.

Prompt effectiveness comparison. We summarize the effectiveness of each prompt category in Table 1, demonstrating that lesion-focused prompts yield the highest clinical validity, while localization prompts exhibit slightly lower consistency due to occasional ambiguity in spatial references.

E. The Biomedical Indicators

The dataset includes 134 structured biomedical indicators encompassing demographic data, laboratory tests, tumor biomarkers, imaging metadata, surgical information, pathological staging, histological findings, and postoperative outcomes are shown in Table 2. These indicators originate from structured EHR and were processed to ensure consistency across patients.

All sensitive identifiers (including patient names, ID numbers, phone numbers, and hospitalization codes) were removed or replaced with anonymized pseudonyms. Variables unrelated to model training (such as historical comorbidities or unused surgery-related entries), are retained for completeness but marked as not used in this study.

Table 2. Full List of 134 Structured Biomedical Variables in the Gastric-X Dataset. De-identified is marked as "De-ID".

Item	Description	Item	Description
Demographics			
Hospital ID (De-ID)	Anonymized hospitalization identifier	Patient Name (De-ID)	Anonymized patient code
Sex	Biological sex	Age (De-ID)	Age at admission
Bed Number (De-ID)	Anonymized bed assignment	Surgery Date (De-ID)	Date of surgery (De-ID)
Imaging ID	CT imaging identifier	CT Description	Radiology description
CBC			
CBC Test Date	Date of CBC test	CBC White Blood Cell Count	White blood cell count
CBC Neutrophil Count	Absolute neutrophil count	CBC Neutrophil Ratio	Neutrophil percentage
CBC Lymphocyte Count	Absolute lymphocyte count	CBC Lymphocyte Ratio	Lymphocyte percentage
CBC Hemoglobin	Hemoglobin concentration	CBC Platelet Count	Platelet count
Biochemistry			
Biochemistry Test Date	Date of biochemistry test	Biochemistry Fasting Glucose	Fasting plasma glucose
Biochemistry Prealbumin	Serum prealbumin	Biochemistry ALT	Alanine aminotransferase
Biochemistry AST	Aspartate aminotransferase	Biochemistry Total Protein	Total serum protein
Biochemistry Albumin	Serum albumin	Biochemistry Total Bilirubin	Total bilirubin
Biochemistry Direct Bilirubin	Direct bilirubin	Biochemistry Creatinine	Serum creatinine
Tumor Markers			
Biochemistry Urea (BUN)	Blood urea nitrogen	Tumor Markers Test Date	Date of tumor marker test
Tumor Markers AFP	Alpha-fetoprotein	Tumor Markers CEA	Carcinoembryonic antigen
Tumor Markers CA125	Cancer antigen 125	Tumor Markers CA724	Cancer antigen 724
Tumor Markers CA199	Cancer antigen 19-9	Past Medical History	Past conditions (not used)
Surgery Details			
Surgery Date	Date of surgery	Resection Range	Extent of resection
Gastrointestinal Reconstruction	Postoperative reconstruction type	Occupation	Patient occupation
Education Level	Highest educational level	Marital Status	Marital status
Ethnicity	Ethnic group	Admission Method	Mode of admission
Insurance Status	Insurance coverage	ID Number (De-ID)	Anonymized ID number
Surgical and Admission Info			
Contact Number (De-ID)	Contact phone	Surgery Admission Date (De-ID)	Admission date for surgery
Surgery Discharge Date (De-ID)	Discharge date	Surgery Hospitalization Cost	Total hospital cost
Admission Temperature	Temperature at admission	Admission Pulse	Pulse rate at admission
Admission Respiration	Respiratory rate	Admission Systolic Pressure	Systolic BP
Admission Diastolic Pressure	Diastolic BP	Height	Height
Weight	Weight	BMI	Body mass index
General Condition	Performance status	Weight Loss	Recent weight loss
Reduced Food Intake	Reduced oral intake	Smoking Status	Smoking history
Drinking Status	Alcohol use	Endoscopy Date (De-ID)	Endoscopy date
Endoscopy Tumor Location	Tumor location	Endoscopy Tumor Size	Tumor size
Endoscopy Gross Type	Gross morphology	Endoscopy Biopsy Pathology	Biopsy pathology
Endoscopy Appearance	Visual findings	Chief Surgeon (De-ID)	Operating surgeon
Tumor Anatomy and Pathology			
Tumor Anatomical Location	Tumor site	Maximum Tumor Diameter	Maximal diameter
Serosal Invasion	Serosal involvement	Gross Tumor Type	Macroscopic type
Linitis Plastica	Linitis plastica presence	Perigastric Lymph Nodes	Perigastric node status
Liver Metastasis	Liver metastasis	Adjacent Organ Invasion	Neighboring organ invasion
Peritoneal Seeding	Peritoneal metastasis	Ascites	Ascites presence
Pathology ID (De-ID)	Specimen ID	Tumor Size (Long Axis)	Long axis size
Tumor Size (Short Axis)	Short axis size	Tumor Size (Height)	Height
Histologic Grade	Differentiation grade	Additional Histologic Type	Additional components
Distance to Proximal Margin	Proximal margin distance	Distance to Distal Margin	Distal margin distance
Specimen Proximal Margin	Proximal margin status	Specimen Distal Margin	Distal margin status
Perineural Invasion	Perineural invasion	Vascular Cancer Thrombus	Vascular invasion
Staging and Lymph Nodes			
T Stage	Pathologic T category	N Stage	Pathologic N category
M Stage	Pathologic M category	Overall Stage	Final stage
Positive Lymph Nodes	Count of positive nodes	Dissected Lymph Nodes	Total dissected nodes
Molecular and IHC			
Ki-67	Proliferation index	HER2 Status	HER2 expression

Continued on next page

Item	Description	Item	Description
CLDN Status	Claudin status	MLH1	MLH1 expression
PMS2	PMS2 expression	MSH2	MSH2 expression
MSH6	MSH6 expression	EBER	EBV RNA (ISH)
PD-L1 Score	PD-L1 score	MMR Status	Mismatch repair status
Complications and Outcomes			
Complication Severity	Clavien–Dindo grade	Secondary Surgery	Reoperation
Complication Occurrence	Any complication	Severe Complication Occurrence	Severe complication
Total Hospital Stay	Total days	Postoperative Stay	Days after surgery
Postoperative Fever	Fever occurrence	Fever Days	Number of fever days
Complication Category	Type of complication	Intervention	Treatment measures
Complication Notes	Additional notes		