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

Sheng Lu[†,*]
Ruijin Hospital

Hao Chen[*]
University of Cambridge

Rui Yin
Nanjing First Hospital

Juyan Ba
Shenzhen University

Yu Zhang
Shanghai Jiao Tong University

Yuanzhe Li[†]
Ruijin Hospital

Abstract

Recent vision-language models (VLMs) have shown strong generalization and multimodal reasoning abilities in natural domains. However, their application to medical diagnosis remains limited by the lack of comprehensive and structured datasets that capture real clinical workflows. To advance the development of VLMs for clinical applications, particularly in gastric cancer, we introduce Gastric-X, a large-scale multimodal benchmark for gastric cancer analysis providing 1.7K cases. Each case in Gastric-X includes paired resting and dynamic CT scans, endoscopic image, a set of structured biochemical indicators, expert-authored diagnostic notes, and bounding box annotations of tumor regions, reflecting realistic clinical conditions. We systematically examine the capability of recent VLMs on five core tasks: Visual Question Answering (VQA), report generation, cross-modal retrieval, disease classification, and lesion localization. These tasks simulate critical stages of clinical workflow, from visual understanding and reasoning to multimodal decision support. Through this evaluation, we aim not only to assess model performance but also to probe the nature of VLM understanding: Can current VLMs meaningfully correlate biochemical signals with spatial tumor features and textual reports? We envision Gastric-X as a step toward aligning machine intelligence with the cognitive and evidential reasoning processes of physicians, and as a resource to inspire the development of next-generation medical VLMs.

1. Introduction

Gastric cancer remains one of the most fatal malignancies worldwide, responsible for over 769,000 deaths in recent global estimates [4, 13]. Its diagnosis requires integrating heterogeneous, multimodal sources of evidence, including multi-phase 3D CT imaging, endoscopic assessment, laboratory tests, and patient history [9, 10, 31, 53, 58]. Among these modalities, CT plays a central role in clinical staging [30], yet evaluations remain subject to substantial inter-observer variability [29]. As the volume and complexity of patient data increase, clinicians face growing cognitive burden, and diagnostic outcomes may vary across institutions and expertise levels [8, 39, 54]. These challenges underscore an urgent need for automated systems capable of jointly perceiving and reasoning across diverse clinical data streams.

In parallel, rapid progress in vision–language models (VLMs) has reshaped our understanding of cross-modal learning in the general domain. Models such as CLIP [49], ALIGN [25], BLIP [33], and Flamingo [1] demonstrate that large-scale alignment of images and text can unlock powerful representations capable of classification, captioning, and visual question answering. These successes have naturally inspired the development of medical VLMs [3, 18, 37, 48]. Yet, the datasets that currently underpin medical VLM research paint a very different picture from real-world clinical practice. Most available benchmarks focus on single-modality 2D radiographs paired with free-text reports, such as MIMIC-CXR [28], CheXpert [23], and PadChest [5]. A few recent efforts have begun to explore volumetric data, e.g., MedVL-CT69K [55] and 3D-RAD [15], but these datasets still primarily emphasize image–report matching and rarely incorporate the richer set of modalities or multi-phase volumes (except MedVL-CT69K) that are indispensable for cancer diagnosis.

This leaves a fundamental gap: *clinical decision-making, especially in oncology, is inherently multimodal.* Radiologists and oncologists must integrate multi-phase 3D imaging, structured laboratory indicators, lesion localization, and clinical narratives to form a coherent diagnosis. Imaging alone provides only a partial view; without datasets that capture this complexity, VLMs often rely on superficial correlations and fail to generalize to real clinical reasoning.

[*]Equal contribution. [†]Correspondence: Lu Sheng and Yuanzhe Li (ls12593@rjh.com.cn, yuanzhe@shuzhiweitech.com).

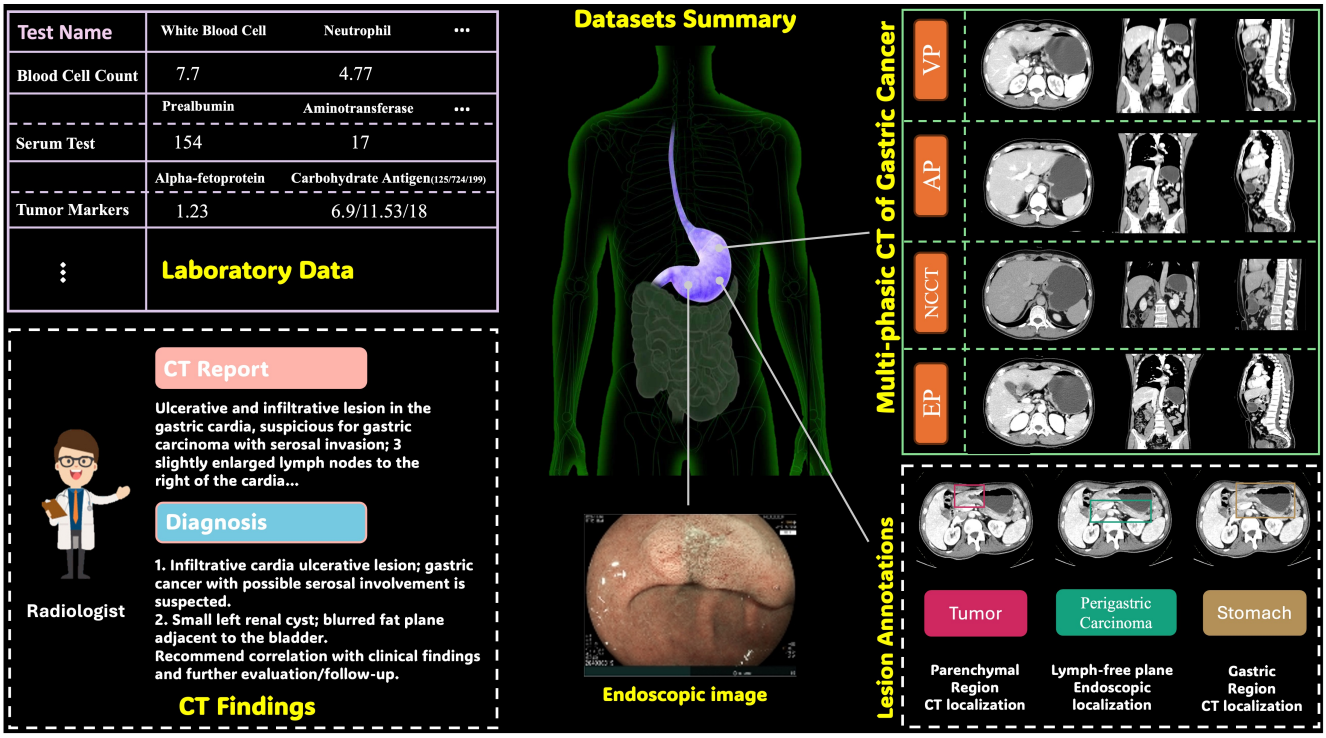


Figure 1. **Overview of the multi-modal information in proposed Gastric-X.** The center panel shows a schematic gastric representation alongside an endoscopic image. The left panel presents examples of structured laboratory data (e.g., blood counts, serum biochemistry, and tumor markers) and clinical textual reports (diagnostic and CT reports) that reflect real-world radiological reasoning. The right panel illustrates multi-phase 3D CT scans (non-contrast, arterial, venous, and equilibrium phases) with multi-class lesion annotations, including tumor, perigastric carcinoma, and stomach regions.

To address this critical gap, we introduce **Gastric-X**, a new multimodal benchmark constructed directly from real-world gastric cancer diagnostic workflows. Gastric-X integrates four key data modalities and annotations essential to clinical decision-making: (1) multi-phase 3D CT scans (non-contrast, arterial, venous, and equilibrium phases) capturing tumor characteristics across enhancement stages [30]; (2) endoscopic images that visualize the tumor; (3) structured patient profiles and biochemical laboratory measurements used in gastric cancer staging [57]; and (4) comprehensive clinical and radiology reports, along with expert-annotated ground truths including disease stages and precise 3D bounding boxes of tumors, which follow the standard clinical workflow for gastric cancer localization. Additionally, the dataset provides curated VQA pairs that probe cross-modal understanding. As shown in Figure 1, the right panel illustrates multi-phase CT imaging and annotated bounding box examples, the left panel presents structured laboratory indicators and clinical reports, and the center highlights the captured endoscopic image. This multimodal set mirrors the information pipeline encountered by radiologists in practice.

Contributions. The main contributions are:

- We introduce **Gastric-X**, a multimodal gastric cancer dataset designed to reflect how clinicians actually reason, by integrating four modalities: multi-phase 3D CT scans, endoscopic images, laboratory measurements, and diagnostic reports, into one unified resource.
- To ground the dataset in real clinical semantics, we provide expert-crafted annotations, including precise 3D bounding boxes for tumors and perigastric lesions, detailed disease stages, and a curated set of VQA pairs that probe cross-modal understanding.
- Building on these components, we establish a comprehensive benchmark of five clinically meaningful tasks, offering a standardized platform for evaluating and advancing multimodal medical VLMs.

2. Related Work

2.1. Vision-Language Models

Vision–Language Models (VLMs) learn joint image–text representations from large-scale data, enabling strong zero-shot generalization [26, 50, 65] beyond traditional supervised learning [17]. The CLIP model [49] established contrastive learning as the foundation for aligning visual

Table 1. **Comprehensive comparison of medical vision–language datasets. Multi-phase:** scans or images captured at different stages or conditions. **Biochemical data:** structured laboratory measurements (e.g., serology results) or Electronic Health Records (EHRs). **Lesion label:** annotations of lesions, including masks or bounding boxes. **Textual Modality:** how text-based labels are provided. The VQA pairs column specifies whether the dataset is primarily designed for visual question answering. The report column indicates whether original diagnostic reports are available.

General Information				Imaging			Annotation & Report		
Dataset	**Release**	**Domain**	**Link**	**Image Modalities**	**#Scans / Images**	**Multi-phase**	**Biochemical Data**	**Lesion Label**	**Textual Modality**
PathVQA [21]	2020	Diverse	link	HP	5.00K Images	✗	✗	✗	VQA Pairs
PadChest [5]	2020	Chest	link	XR	160K Images	✗	✗	✓	Reports
SLAKE [40]	2022	Diverse	link	CT, MR, XR	642 Items	✗	✗	✓	VQA Pairs
Merlin [3]	2024	Abdomen	link	CT	25.5K Scans	✗	✓	✗	Reports
MIMIC-CXR v2 [27]	2024	Chest	link	XR	377K Images	✗	✓	✗	Reports
CT-RATE [19]	2024	Chest	link	CT	25.7K Scans	✗	✗	✗	Reports
GEMeX [41]	2025	Chest	link	XR	151K Images	✗	✗	✗	VQA Pairs
MedVL-CT69K [7]	2025	Diverse	–	CT	272K Scans	✓	✗	✗	Reports
3D-RAD [15]	2025	Diverse	link	CT	16.1K Scans	✗	✗	✗	VQA Pairs
Gastric-X	2025	Gastric	–	CT	7.1K Scans 1.7K Images	✓	✓	✓	Reports

Notes. Diverse: includes multiple organs. CT: Computed Tomography; HP: Histopathology; MR: Magnetic Resonance; XR: X-ray.

and linguistic spaces. Building on this, later works introduced multi-objective training [56, 67] and unified architectures [24, 60], extending VLM capabilities to dense prediction tasks such as object detection [35, 66]. For downstream adaptation, lightweight techniques such as prompt tuning [73, 74] and feature adaptation [16] efficiently tailor pretrained models to new tasks, enabling VLMs to serve as versatile foundations for multimodal reasoning.

2.2. VLMs in Medical Diagnosis

Inspired by the success of general-domain VLMs, medical adaptations have sought to extend them to support clinical diagnosis. Early studies primarily aligned 2D X-ray images with radiology reports through contrastive objectives [11, 12, 59, 72]. Subsequent work refined these representations via fine-grained region–text correspondence [22, 48, 61] and integration of structured medical knowledge [36, 42, 64, 69, 71]. However, most efforts remain limited to 2D imagery, with only recent attempts exploring 3D visual–language learning from computed tomography (CT) volumes [3, 6, 18, 37]. A further challenge lies in the reliance on single-volume inputs, while real-world diagnosis often requires multiple imaging sequences. For instance, 4 MRI sequences for brain tumors or 2 (ideally 4) CT volumes for gastric cancer. These gaps highlight the need for domain-specific 3D VLM frameworks that better reflect clinical diagnostic practice.

2.3. Multi-Modal Datasets for VLM Diagnosis

Progress in medical VLMs is closely tied to the availability of high-quality, multi-modal datasets. Foundational resources such as TCIA [14] and TCGA [62] have provided extensive data that underpin cancer research. In radiology, dataset development has rapidly evolved, from early 2D efforts linking images to free-text reports, such as MIMIC-CXR [2] and PadChest [5], to vision–language benchmarks with structured supervision like SLAKE [40] and GEMeX [41]. More recently, 3D CT datasets including MedVL-CT69K [55] and 3D-RAD [15] have shifted attention toward volumetric reasoning and multi-modal integration. While these developments mark important progress toward holistic medical understanding, multi-modal datasets dedicated to cancer diagnosis such as gastric cancer remain scarce, limiting the advancement of specialized 3D VLMs in this critical domain.

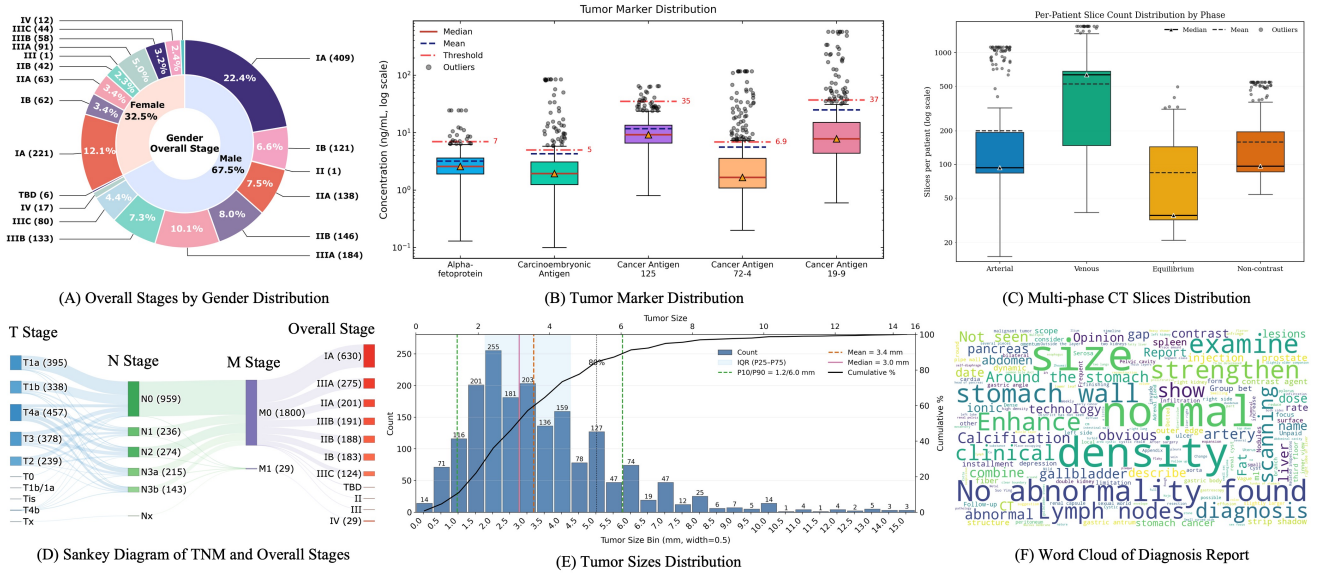
3. Gastric-X: Design and Construction

In this section, we introduce Gastric-X, a clinically inspired multi-modal dataset designed to bridge real-world diagnostic workflows and vision–language modeling. We first describe its clinical motivation and guiding design principles, followed by a detailed overview of its multimodal composition and associated reasoning tasks.

3.1. Overview and Background

Gastric-X is a multimodal gastric cancer dataset specifically curated for vision–language model research. Its design is motivated by how clinicians diagnose gastric cancer in practice. Clinically, diagnosis is never based on a single modality. Radiologists analyze multiple CT phases, gastroenterologists interpret endoscopic findings. These imaging assessments are integrated with biochemical tests and laboratory results to form a comprehensive diagnostic conclusion.

This complex multi-model reliance process inspired the creation of our *Gastric-X*: a dataset that mirrors real diagnostic reasoning by aligning these heterogeneous data sources within a unified framework. Each patient record in Gastric-X includes quad-phase CT scans, endoscopic im-



(A) Overall Stages by Gender Distribution (B) Tumor Marker Distribution (C) Multi-phase CT Slices Distribution

(D) Sankey Diagram of TNM and Overall Stages (E) Tumor Sizes Distribution (F) Word Cloud of Diagnosis Report

Figure 2. **Gastric-X Overview.** (A) Overall stage distribution by gender (67.5% male, 32.5% female). (B) Distribution of 5 tumor markers on a logarithmic scale. (C) Overall distribution of CT slice lengths across 4 different phases. (D) Sankey diagram illustrating the transitions between T, N, and M stages and corresponding overall stages. (E) Histogram of tumor sizes with a cumulative percentage curve. (F) Word cloud summarizing the most frequent terms from radiology reports. Zoom in for better visualization.

ages, biochemical indicators, expert bounding box annotations of lesions, detailed imaging and endoscopic reports, and final diagnostic summaries. To facilitate advanced reasoning tasks, Gastric-X is further enriched with Visual Question Answering (VQA) annotations, making it a comprehensive benchmark for clinical VLM development. The overview of dataset distribution can be seen in Figure 2.

3.2. Multi-modal Data

Our dataset is inherently multi-modal, encompassing diverse types of medical data that reflect how clinicians integrate information from multiple sources during diagnosis.

Imaging Modality. The foundation of *Gastric-X* is its diverse and complementary imaging data. We begin with the CT modality, collected across four key phases: arterial, venous, equilibrium, and non-contrast. Each phase offers a unique view of vascular perfusion and tissue enhancement, and together they capture the dynamic passage of contrast through the gastric wall. In clinical practice, radiologists often integrate these dynamic differences to form a holistic understanding of disease progression. *Gastric-X* reflects this process in data form.

Complementing CT, *Gastric-X* also includes endoscopic images, which provide high-resolution, color-rich views of the gastric mucosa. These images expose fine-grained textures, color variations, and microvascular structures invisible to CT. Subtle irregularities, such as fold distortion, erosion, or discoloration, often signal the earliest stages of malignancy and help pinpoint tumor location and stage.

Gastric-X comprises 7.1K CT scans (a total of 83.48K slices) and 1.7K endoscopic images. We unify these complementary views to support a holistic understanding of gastric cancer through multi-modal imaging. The distribution of multi-phase CT slices is shown in Figure 2C.

Stages and Lesion Annotation. Gastric cancer is clinically staged using the TNM system, which characterizes disease progression by assessing the primary tumor (T), regional lymph nodes (N), and distant metastasis (M). This framework guides diagnosis, treatment planning, and prognosis estimation, ultimately determining the overall stage. The distribution of overall stages by gender is shown in Figure 2A. The transition from TNM components to overall stages is illustrated in Figure 2D, and the tumor size distribution is presented in Figure 2E.

For each CT phase, *Gastric-X* provides detailed 3D bounding box (BBox) annotations covering both tumor lesions and relevant organs. The annotations are organized into three hierarchical levels to capture different diagnostic focuses: the tumor core, the regional lymph nodes, and the entire stomach region. In total, each CT study includes three BBoxes per phase across four phases and 1.74K patients, resulting in a rich and comprehensive annotation set of 21,408 BBoxes for multi-scale lesion analysis.

Biochemical Data. Beyond imaging, the diagnostic reasoning of clinicians often turns to the language of biochemical test, signals that quietly reflect internal pathology. *Gastric-X* captures this dimension through 11 serum biochemistry

indicators (*e.g.* liver and renal function markers) and 5 key tumor markers commonly used in gastric cancer screening. To complement these laboratory measures, we further include comprehensive electronic health records (EHR) detailing surgical procedures, medication history, and treatment progress across 134 structured items. The tumor marker distribution is shown in Figure 2B.

Reports. Beyond imaging, *Gastric-X* incorporates rich clinical reports that mirror the reasoning chain of real-world diagnosis. Each patient record is paired with three complementary types of reports: (1) the CT report, describing lesion morphology, enhancement patterns, and staging impressions; (2) the endoscopic report, detailing mucosal appearance, color, and biopsy findings; and (3) the diagnosis report, which summarizes pathological confirmation and clinical outcomes. The word cloud of diagnosis reports is shown in Figure 2F.

Building upon these reports, we construct 26,760 visual question–answer (VQA) pairs, designed to transform narrative clinical observations into structured reasoning tasks.

3.3. Comparison to Other Datasets

As summarized in Table 1, existing medical vision–language datasets primarily focus on static imaging modalities (e.g., X-ray, CT, pathology slides) paired with diagnostic reports or VQA annotations. While datasets such as PathVQA [21], SLAKE [40], and GEMeX [41] have advanced multimodal research, they lack dynamic information from multi-phase imaging and structured clinical measurements.

More recent datasets, including Merlin [3] and MIMIC-CXR [27], begin to incorporate biochemical data. Structured EHR components such as serology results offer quantitative insights into physiological states not captured by imaging alone. For gastric cancer specifically, TCGA-STAD [46] provides 46 subjects with CT scans and accompanying clinical, genomic, and histopathology data, but remains limited in scale.

In contrast, *Gastric-X* introduces a previously missing multimodal configuration: multi-phase 3D CT, endoscopy, structured biochemistry, lesion annotations, and clinical text, all aligned at the patient level. No existing dataset offers such a comprehensive setting. This design provides the multimodal supervision necessary for realistic clinical reasoning and establishes a new standard for holistic medical understanding.

4. Adapting VLMs to Our Multi-Modal Data

To evaluate how current vision–language models (VLMs) handle our multi-model medical data, we adapt both general-purpose and medical-specific models to Gastric-X. The symbolic adaptation and the tasks can be seen in Figure 3.

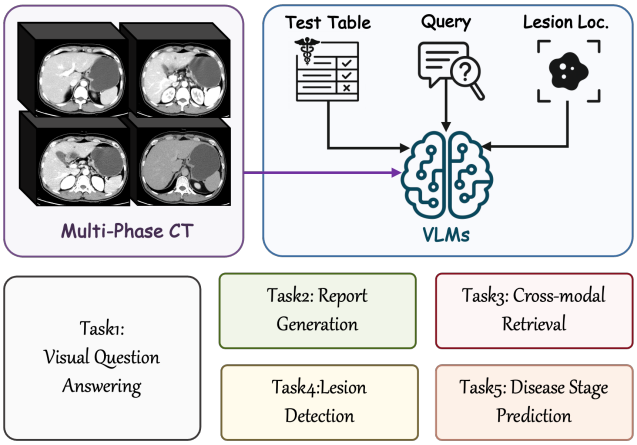


Figure 3. **VLM adaptation.** In adapting VLMs to our dataset, the visual encoder incorporates multi-phase CT inputs, while test tables, textual queries, and lesion-localization cues serve as complementary multimodal inputs guiding the model's diagnostic reasoning. The VLMs must effectively adapt to these diverse input modalities to better capture and perform the targeted clinical tasks.

Model Selection. We select LLaVA-1.5-7B [43], BLIP-2 [34], and X^2-VLM [68] as representative general VLMs, and LLaVA-Med v1.5 [32], Med-Flamingo [47], and Med-VInT [70] as medical-domain models.

CT Adaptation. Our dataset comprises multi-phase 3D CT volumes, whereas most VLMs are designed for 2D or single-image inputs. For LLaVA-1.5 [43] and BLIP-2 [34], which natively support multi-channel image inputs, we concatenate CT slices across phases to form multi-channel representations. For X^2-VLM [68], whose architecture is 2D, we replace its vision encoder with a 3D Swin Transformer [44, 45] and its text encoder with Med-BERT [51], both initialized with pre-trained weights. The adapted model is referred to as X^2-VLM-Med.

For medical-specific models, Med-Flamingo and Med-VInT already support volumetric inputs in the form of sequential 2D (pseudo-3D) processing and therefore require only minimal modification. In contrast, LLaVA-Med natively accepts only 2D inputs, so we adopt the same strategy used for X^2-VLM [68] by replacing its vision encoder with a Swin Transformer [45] to enable effective 3D volume processing.

Retrieval Extension. Some models lack built-in retrieval capability. We add a lightweight retrieval head to enable bidirectional image–text matching, allowing consistent evaluation across all models on cross-modal retrieval tasks.

Addition modalities usage. Clinical diagnosis rarely relies on imaging alone; clinicians naturally integrate visual cues with structured biomedical measurements. To mirror this workflow with minimal architectural changes, we introduce two lightweight but impactful sources of auxiliary informa-

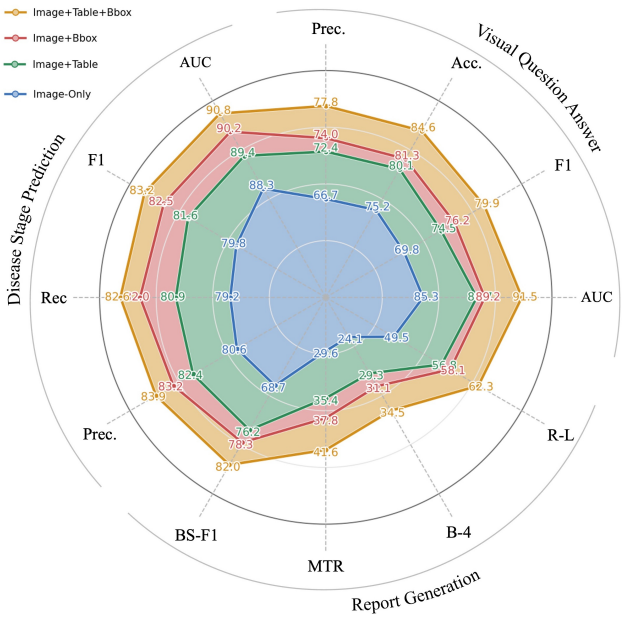


Figure 4. Radar plot comparing multimodal configurations across three medical vision-language tasks. The "Image+Table+Bbox" configuration achieves the highest overall performance across all evaluation metrics.

tion: bounding-box cues and biomedical test tables, to support the VLMs. The bounding boxes are rendered directly on the CT slices as colored overlays, serving as soft spatial priors that nudge the VLM's attention toward clinically relevant regions without altering the visual encoder.

In contrast, biomedical test tables present a unique challenge: they contain dozens of measurements, many of which are irrelevant or fall within normal ranges. Rather than feeding the entire table, we mimic the behavior of clinicians, who first look for abnormal values, meaning values that exceed physiological thresholds, and reason from there. We extract only these abnormal entries and convert them into concise textual descriptors that include the test name, its measured value, and the ratio by which it exceeds the normal limit.

5. Experiments

Implementation Details. All models are fine-tuned using the AdamW optimizer with a batch size of 32 and a learning rate of 5e-5, following a linear warm-up and decay schedule. All models are trained on a single NVIDIA RTX 3090 GPU. The dataset is split at the patient level into training, validation, and test sets (70/15/15). All models are initialized from the X2-VLM checkpoint and trained independently for each task. We optimize models using AdamW with a learning rate of 5×10^{-5}, weight decay 0.01, and 10% warm-up schedule. The text encoder is trained with a $2\times$learning rate relative to the visual encoder.

Multi-Modal Input Settings. In clinical practice, multiple modalities are often examined together to support diagnosis. To emulate this process and evaluate the contribution of each modality, we design four input schemas: (1) *Image Only*, using visual information alone; (2) *Image + Table*, combining images with biomedical test results; (3) *Image + BBox*, incorporating both images and bounding box annotations; and (4) *Image + Table + BBox*, integrating all available modalities for a comprehensive diagnostic setting.

Task Settings. To comprehensively evaluate multi-modal understanding, we design five challenges within our dataset: report generation, visual question answering, cross-modal retrieval, phase classification, and lesion detection.

For all task result tables, the best and second-best results are indicated in **bold** and underlined, respectively.

A summary of the model performance across all configurations in the three main tasks is illustrated in Figure 4. The results are reported using X^2-VLM-Med [68], demonstrating consistent improvements when incorporating multimodal information. Notably, the *Image+Table+Bbox* configuration achieves the most balanced and superior results across all evaluation metrics.

The detailed configurations and results for each task are presented in the following subsections.

5.1. Visual Question Answering

As shown in Table 3a, across all four evaluation settings, a consistent upward trend is observed in model performance as the input conditions become progressively richer and complex. Specifically, all models exhibit improvements in Precision, Accuracy, F1, and Area Under the Curve (AUC) from the image-only setting to the fully integrated one, indicating that exposure to more informative or balanced input–target combinations substantially enhances visual–language reasoning.

Among the compared methods, X^2-VLM-Med [68] achieves the best overall performance across all metrics and settings, demonstrating superior generalization and multimodal alignment capabilities. Its AUC steadily increases from 85.3% (Image Only) to 91.5% (Image + Table + BBox), reflecting robust discriminative power under varying configurations. Med-Flamingo [47] consistently ranks second, highlighting its strong medical-domain adaptation, while MedVInT [70] closely follows with competitive gains across all metrics.

5.2. Report Generation

We evaluate report generation quality using four complementary metrics: ROUGE-L (R-L), BLEU-4 (B-4), METEOR (MTR), and BERTScore F1 (BS-F1), to cover lexical overlap, fluency, and semantic fidelity between generated

Table 2. Performance across modalities for Vision Question Answer and report generation. We evaluate four input-modality settings, including Image Only and combinations that add Table, Bounding Box (BBox), or both. The best and second-best results in each column are shown in **bold** and underlined, respectively.

Model	Image Only				Image + Table				Image + BBox				Image + Table + BBox			
	Prec. ↑	Acc. ↑	F1 ↑	AUC (%) ↑	Prec. ↑	Acc. ↑	F1 ↑	AUC (%) ↑	Prec. ↑	Acc. ↑	F1 ↑	AUC (%) ↑	Prec. ↑	Acc. ↑	F1 ↑	AUC (%) ↑
LLaVA-1.5-7B [43]	42.5	53.1	46.2	67.8	48.3	59.7	52.4	72.1	50.1	61.0	54.3	73.0	55.2	66.5	59.1	77.8
LLaVA-Med v1.5 [32]	56.3	64.7	59.9	78.4	62.5	71.8	65.7	82.0	63.8	72.9	66.9	82.6	67.4	76.1	69.8	85.0
Med-Flamingo [47]	60.8	68.9	63.9	80.5	66.9	74.8	69.7	84.2	68.1	75.4	70.9	84.8	71.0	78.9	73.5	86.5
BLIP-2 [34]	50.2	59.4	53.7	73.1	56.0	64.8	58.9	77.0	58.0	66.4	60.7	78.2	61.3	69.7	63.9	80.1
MedVInT [70]	58.1	66.0	60.9	79.0	63.4	71.5	65.9	82.5	65.0	72.3	67.4	83.0	68.0	75.8	70.1	85.4
X^2-VLM-Med [68]	**66.7**	**75.2**	**69.8**	**85.3**	**72.4**	**80.1**	**74.5**	**88.7**	**74.0**	**81.3**	**76.2**	**89.2**	**77.8**	**84.6**	**79.9**	**91.5**

(a) **Visual Question Answering.** Prec.: Precision, Acc.: Accuracy, AUC: Area Under the ROC Curve.

Model	Image Only				Image + Table				Image + BBox				Image + Table + BBox			
	R-L ↑	B-4 ↑	MTR ↑	BS-F1 ↑	R-L ↑	B-4 ↑	MTR ↑	BS-F1 ↑	R-L ↑	B-4 ↑	MTR ↑	BS-F1 ↑	R-L ↑	B-4 ↑	MTR ↑	BS-F1 ↑
LLaVA-1.5-7B [43]	28.4	9.2	14.1	45.3	34.2	12.1	17.6	51.2	36.0	13.5	18.9	53.1	40.1	16.3	21.7	57.8
LLaVA-Med v1.5 [32]	35.7	13.8	18.9	53.6	42.0	17.2	23.4	60.0	44.1	18.7	25.1	62.6	48.7	21.9	28.4	66.5
Med-Flamingo [47]	42.1	18.6	23.4	60.2	49.0	22.4	28.2	67.0	51.2	24.1	30.1	69.4	55.6	27.8	34.6	73.1
BLIP-2 [34]	36.8	15.2	19.7	55.1	43.5	19.0	25.0	61.9	45.6	20.6	26.8	64.0	50.2	23.7	30.5	67.7
MedVInT [70]	38.9	16.9	21.0	57.3	45.1	20.8	26.9	63.5	47.2	22.3	28.7	65.9	52.0	25.5	32.4	69.8
X^2-VLM-Med [68]	**49.5**	**24.1**	**29.6**	**68.7**	**56.8**	**29.3**	**35.4**	**76.2**	**58.1**	**31.1**	**37.8**	**78.3**	**62.3**	**34.5**	**41.6**	**82.0**

(b) **Report Generation.** Evaluation metrics include ROUGE-L (R-L), BLEU-4 (B-4), METEOR (MTR), and BERTScore F1 (BS-F1).

and reference reports.

As shown in Table 3b, all models exhibit a consistent upward trend across the four experimental settings, mirroring the pattern observed in Table 3a. X^2-VLM-Med [68] achieves the best overall performance, with BERTScore F1 improving from 68.7 in Image Only to 82.0 in Image + Table + BBox, highlighting its strong cross-modal reasoning and text generation capabilities.

5.3. Cross-modal Retrieval

Table 4 presents image-only cross-modal retrieval results for both Image→Text and Text→Image, evaluated using Recall at K (Recall@K), Median Rank (MedR), Mean Rank (MnR), and Mean Average Precision (mAP).

Across all metrics, X^2-VLM-Med [68] achieves the best performance, reaching an R@1 of 48.9% for Image→Text and 47.5% for Text→Image, substantially outperforming other models. Med-Flamingo [47] consistently ranks second, reflecting effective multimodal reasoning within the medical domain.

MedVInT [70] and BLIP-2 [34] deliver competitive results but remain behind the top two models, while LLaVA-Med v1.5 [32] and LLaVA-1.5-7B [43] lag further due to limited retrieval-specific adaptation.

5.4. Disease Stage Classification

Table 5 presents the results for disease stage classification across all evaluated models. We observe a consistent performance hierarchy across input configurations, broadly aligned with both model capacity and multimodal alignment strength. Conventional convolutional architectures such as ResNet-50 [20] lag considerably behind transformer-based approaches, reflecting their limited ability to capture long-range and cross-modal dependencies. In contrast, the Swin Transformer [44] achieves notable gains over ResNet baselines, confirming that hierarchical self-attention effectively models localized medical features and spatial-scale variations.

Among multimodal encoders, MedVInT [70] and LLaVA-Med v1.5 [32] demonstrate moderate improvements, particularly when incorporating table cues. However, their performance saturates as the modality complexity increases (Image + Table + BBox).

In contrast, X^2-VLM-Med [68] consistently achieves the strongest results across all metrics and input configurations, outperforming the second-best model by a clear margin. The gain is most pronounced in AUC (+1.6 over Swin on Image + Table + BBox).

5.5. Lesion Detection

Table 6 summarizes lesion detection performance under COCO evaluation settings [38], reporting COCO-style AP and F1. Faster R-CNN [52] serves as a strong convolutional baseline but shows moderate precision, with AP@0.5 = 64.1 and localization accuracy of 70.4.

The Swin Transformer [63] significantly improves across all metrics, benefiting from hierarchical self-attention that captures multi-scale lesion structures. MedVInT [70] further boosts performance, attaining the highest AP@0.5 of 72.1 and a mean AP of 50.2, reflecting superior spatial reasoning through multimodal medical pretraining.

LLaVA-Med v1.5 [43] maintains strong overall detection quality, especially in F1@0.5, suggesting effective vision–language grounding despite limited domain adapta-

Table 4. **Cross-modal retrieval.** This table reports retrieval results in both Image-to-Text and Text-to-Image directions. Metrics include Recall@K (%, higher is better), Median Rank and Mean Rank (lower is better), and mean Average Precision (mAP). **Bold** and underlined numbers denote the best and second-best performance in each column.

Model	Image → Text						Text → Image					
	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓	MnR ↓	mAP ↑	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓	MnR ↓	mAP ↑
LLaVA-1.5-7B [43]	24.3	52.7	63.1	11.2	28.4	38.5	22.1	50.3	61.7	12.4	29.8	36.9
LLaVA-Med v1.5 [32]	35.6	68.9	78.3	7.5	19.6	52.7	33.8	65.2	76.1	8.1	21.4	50.3
Med-Flamingo [47]	42.8	74.1	83.5	6.2	17.3	57.9	41.5	72.6	82.8	6.6	18.4	56.8
BLIP-2 [34]	39.7	70.5	80.9	6.9	18.1	55.3	37.6	68.9	79.2	7.4	19.7	53.2
MedVInT [70]	40.1	73.6	82.4	6.4	17.1	56.4	39.4	71.8	81.2	6.8	17.9	55.1
X^2-VLM-Med [68]	**48.9**	**80.7**	**88.2**	**4.9**	**13.5**	**63.1**	**47.5**	**79.3**	**87.4**	**5.2**	**14.1**	**61.7**

Table 5. **Disease stage classification.** For each configuration, we evaluate Precision, Recall, F1 score, and Area Under (AUC). These results show how integrating multimodal cues and medical-aware pretraining benefits fine-grained disease staging. **Bold** and underlined numbers indicate the best and second-best performance in each column.

Model	Image Only				Image + Table				Image + BBox				Image + Table + BBox			
	Prec. ↑	Rec. ↑	F1 ↑	AUC ↑	Prec. ↑	Rec. ↑	F1 ↑	AUC ↑	Prec. ↑	Rec. ↑	F1 ↑	AUC ↑	Prec. ↑	Rec. ↑	F1 ↑	AUC ↑
ResNet-50 [20]	72.8	70.6	71.6	80.3	74.1	72.0	73.0	81.2	75.9	73.7	74.7	82.1	76.4	74.8	75.5	82.7
Swin Transformer [44]	79.5	78.0	78.7	87.1	80.4	79.2	79.8	88.0	82.1	80.9	81.5	89.2	83.4	82.0	82.7	90.1
MedVInT [70]	77.1	76.2	76.6	85.5	81.0	79.8	80.4	87.6	80.6	79.4	80.0	88.3	81.2	80.1	80.6	88.7
LLaVA-Med v1.5 [32]	75.8	74.3	75.0	84.0	78.0	76.6	77.3	85.3	78.8	77.5	78.1	86.1	79.5	78.0	78.7	86.6
X^2-VLM-Med [68]	**80.6**	**79.2**	**79.8**	**88.3**	**82.4**	**80.9**	**81.6**	**89.4**	**83.2**	**82.0**	**82.5**	**90.2**	**83.9**	**82.6**	**83.2**	**90.8**

Table 6. **Lesion detection**. Metrics are Average Precision (AP) and F1. *mAP* denotes mean AP averaged over IoU thresholds from 0.50 to 0.95 (step 0.05), and localization accuracy (Loc. Acc.) is computed at IoU = 0.5.

Model	AP@0.5	AP@0.75	F1@0.5	mAP	Loc. Acc.
FRCNN [52]	64.1	48.7	61.0	43.2	70.4
Swin Transformer [63]	70.5	53.9	66.7	48.9	77.1
MedVInT [70]	**72.1**	55.1	67.8	50.2	78.5
LLaVA-Med v1.5 [43]	68.3	51.7	67.9	46.8	75.2
X^2-VLM-Med [68]	70.4	**56.4**	**68.1**	**51.5**	**79.6**

Metric@IoU indicates computation at a given IoU threshold.

tion. The X^2-VLM-Med [68] achieves the best results on most evaluation metrics, including AP@0.75, F1@0.5, mAP, and localization accuracy, outperforming all baselines by a consistent margin.

6. Ethical Concern and Publication

All data were collected by first-line clinicians under strict institutional ethical approval and in full compliance with relevant privacy and data-protection regulations. After collection, all samples were fully de-identified and manually checked to ensure that no personally identifiable information remained.

The annotations were created by experienced clinicians with domain expertise and were cross-checked for consistency. We aim to ensure that the clinicians' qualifications and the review process provide a solid basis for the reliability and clinical validity of the annotations, which may benefit future research and model development.

The dataset is well-structured, has obtained Institutional Review Board (IRB) approval, and is planned to be made publicly available after the paper is accepted. The release will follow institutional data-sharing policies, and no data that restrict public or research redistribution will be included. We plan to host a small subset of samples on Hugging Face for demonstration purposes, while the complete dataset will be distributed through our project webpage after publication. Access to the dataset will require signing a consent form prior to distribution. The dataset is intended to be released under the CC BY-NC-ND 4.0 license.

7. Conclusion

We presented **Gastric-X**, a comprehensive multimodal benchmark designed to advance vision–language research in gastric cancer analysis. The dataset is carefully curated from real clinical workflows, integrating multi-phase 3D CT scans, endoscopic images, biochemical indicators, and clinical reports with bounding box and disease stage annotations. Gastric-X provides a unified platform encompassing five core tasks, including visual question answering, report generation, cross-modal retrieval, disease stage classification, and lesion detection, offering a holistic evaluation of multimodal reasoning in medical AI. We will publicly release the dataset and accompanying experimental code to support reproducibility and community development. We envision Gastric-X as a standard reference for developing robust and clinically aligned vision–language models in healthcare.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1

[2] Elisa Baratella, Cristina Marrocchio, Alessandro Marco Bozzato, Erik Roman-Pognuz, and Maria Assunta Cova. Chest x-ray in intensive care unit patients: what there is to know about thoracic devices. *Diagnostic and Interventional Radiology*, 27(5):633, 2021. 3

[3] Louis Blankemeier, Joseph Paul Cohen, Ashwin Kumar, Dave Van Veen, Syed Jamal Safdar Gardezi, Magdalini Paschali, Zhihong Chen, Jean-Benoit Delbrouck, Eduardo Reis, Cesar Truyts, et al. Merlin: A vision language foundation model for 3d computed tomography. *Research Square*, pages rs–3, 2024. 1, 3, 5

[4] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424, 2018. 1

[5] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020. 1, 3

[6] Weiwei Cao, Jianpeng Zhang, Yingda Xia, Tony CW Mok, Zi Li, Xianghua Ye, Le Lu, Jian Zheng, Yuxing Tang, and Ling Zhang. Bootstrapping chest ct image understanding by distilling knowledge from x-ray expert models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11238–11247, 2024. 3

[7] Weiwei Cao, Jianpeng Zhang, Zhongyi Shui, Sinuo Wang, Zeli Chen, Xi Li, Le Lu, Xianghua Ye, Qi Zhang, Tingbo Liang, et al. Boosting vision semantic density with anatomy normality modeling for medical vision-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23041–23050, 2025. 3

[8] Hao Chen, Hongrun Zhang, U Wang Chan, Rui Yin, Xiaofei Wang, and Chao Li. Domain game: Disentangle anatomical feature for single domain generalized segmentation. In *International Workshop on Computational Mathematics Modeling in Cancer Analysis*, pages 41–51. Springer Nature Switzerland Cham, 2024. 1

[9] Hao Chen, Rui Yin, Yifan Chen, Qi Chen, and Chao Li. Learning patient-specific disease dynamics with latent flow matching for longitudinal imaging generation. *ICLR 2026*, 2025. 1

[10] Yifan Chen, Fei Yin, Hao Chen, Jia Wu, and Chao Li. Pmpbench: A paired multi-modal pan-cancer benchmark for medical image synthesis. *arXiv preprint arXiv:2601.15884*, 2026. 1

[11] Zhihong Chen, Guanbin Li, and Xiang Wan. Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge. In *Proceedings of the 30th ACM international conference on multimedia*, pages 5152–5161, 2022. 3

[12] Pujin Cheng, Li Lin, Junyan Lyu, Yijin Huang, Wenhan Luo, and Xiaoying Tang. Prior: Prototype representation joint learning from medical images and reports. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 21361–21371, 2023. 3

[13] Bhupender S Chhikara, Keykavous Parang, et al. Global cancer statistics 2022: the trends projection analysis. *Chemical biology letters*, 10(1):451–451, 2023. 1

[14] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26(6):1045–1057, 2013. 3

[15] Xiaotang Gai, Jiaxiang Liu, Yichen Li, Zijie Meng, Jian Wu, and Zuozhu Liu. 3d-rad: A comprehensive 3d radiology med-vqa dataset with multi-temporal analysis and diverse diagnostic tasks. *arXiv preprint arXiv:2506.11147*, 2025. 1, 3

[16] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2): 581–595, 2024. 3

[17] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 2

[18] Ibrahim Ethem Hamamci, Sezgin Er, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Muhammed Furkan Dasdelen, Bastian Wittmann, Enis Simsar, Mehmet Simsar, et al. A foundation model utilizing chest ct volumes and radiology reports for supervised-level zero-shot detection of abnormalities. *CoRR*, 2024. 1, 3

[19] Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Doga, Omer Faruk Durugol, Weicheng Dai, Murong Xu, et al. Developing generalist foundation models from a multimodal dataset for 3d computed tomography. *arXiv preprint arXiv:2403.17834*, 2024. 3

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7, 8

[21] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020. 3, 5

[22] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. 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. 3

[23] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert:

A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, pages 590–597, 2019. 1

[24] Jiho Jang, Chaerin Kong, Donghyeon Jeon, Seonhoon Kim, and Nojun Kwak. Unifying vision-language representation space with single-tower transformer. In *Proceedings of the AAAI conference on artificial intelligence*, pages 980–988, 2023. 3

[25] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1

[26] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2

[27] Alistair Johnson, Tom Pollard, Roger Mark, Seth Berkowitz, and Steven Horng. Mimic-cxr database (version 2.1.0), 2024. RRID:SCR_007345. 3, 5

[28] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019. 1

[29] Jung Hoon Kim, Hyo Won Eun, Jae Ho Choi, Seong Sook Hong, Weechang Kang, and Yong Ho Auh. Diagnostic performance of virtual gastroscopy using mdct in early gastric cancer compared with 2d axial ct: focusing on interobserver variation. *American Journal of Roentgenology*, 189(2):299–305, 2007. 1

[30] Robert Michael Kwee and Thomas Christian Kwee. Imaging in local staging of gastric cancer: a systematic review. *Journal of clinical oncology*, 25(15):2107–2116, 2007. 1, 2

[31] Robert Michael Kwee and Thomas Christian Kwee. Imaging in local staging of gastric cancer: a systematic review. *Journal of clinical oncology*, 25(15):2107–2116, 2007. 1

[32] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023. 5, 7, 8

[33] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 1

[34] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 5, 7, 8

[35] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10975, 2022. 3

[36] Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. 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. 3

[37] Jingyang Lin, Yingda Xia, Jianpeng Zhang, Ke Yan, Le Lu, Jiebo Luo, and Ling Zhang. Ct-glip: 3d grounded language-image pretraining with ct scans and radiology reports for full-body scenarios. *arXiv preprint arXiv:2404.15272*, 2024. 1, 3

[38] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014. 7

[39] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017. 1

[40] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE, 2021. 3, 5

[41] Bo Liu, Ke Zou, Li-Ming Zhan, Zexin Lu, Xiaoyu Dong, Yidi Chen, Chengqiang Xie, Jiannong Cao, Xiao-Ming Wu, and Huazhu Fu. Gemex: A large-scale, groundable, and explainable medical vqa benchmark for chest x-ray diagnosis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21310–21320, 2025. 3, 5

[42] Chang Liu, Yuanhe Tian, Weidong Chen, Yan Song, and Yongdong Zhang. Bootstrapping large language models for radiology report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18635–18643, 2024. 3

[43] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024. 5, 7, 8

[44] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 5, 7, 8

[45] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 5

[46] F. R. Lucchesi and N. D. Aredes. The cancer genome atlas stomach adenocarcinoma collection (tcga-stad). https://doi.org/10.7937/K9/TCIA.2016.GDHL9KIM, 2016. Data set, The Cancer Imaging Archive. 5

[47] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023. 5, 6, 7, 8

[48] Philip Müller, Georgios Kaissis, Congyu Zou, and Daniel Rueckert. Joint learning of localized representations from medical images and reports. In *European conference on computer vision*, pages 685–701. Springer, 2022. 1, 3

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[51] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86, 2021. 5

[52] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 7, 8

[53] Rachel E Sexton, Mohammed Najeeb Al Hallak, Maria Diab, and Asfar S Azmi. Gastric cancer: a comprehensive review of current and future treatment strategies. *Cancer and Metastasis Reviews*, 39(4):1179–1203, 2020. 1

[54] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19(1):221–248, 2017. 1

[55] Zhongyi Shui, Jianpeng Zhang, Weiwei Cao, Sinuo Wang, Ruizhe Guo, Le Lu, Lin Yang, Xianghua Ye, Tingbo Liang, Qi Zhang, et al. Large-scale and fine-grained vision-language pre-training for enhanced ct image understanding. *arXiv preprint arXiv:2501.14548*, 2025. 1, 3

[56] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15638–15650, 2022. 3

[57] EC Smyth, M Verheij, W Allum, D Cunningham, A Cervantes, D Arnold, ESMO Guidelines Committee, et al. Gastric cancer: Esmo clinical practice guidelines for diagnosis, treatment and follow-up. *Annals of oncology*, 27:v38–v49, 2016. 2

[58] EC Smyth, M Verheij, W Allum, D Cunningham, A Cervantes, D Arnold, ESMO Guidelines Committee, et al. Gastric cancer: Esmo clinical practice guidelines for diagnosis, treatment and follow-up. *Annals of oncology*, 27:v38–v49, 2016. 1

[59] Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature biomedical engineering*, 6(12): 1399–1406, 2022. 3

[60] Michael Tschannen, Basil Mustafa, and Neil Houlsby. Image-andlanguage understanding from pixels only. *arXiv preprint arXiv:2212.08045*, 3, 2022. 3

[61] Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Advances in neural information processing systems*, 35: 33536–33549, 2022. 3

[62] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013. 3

[63] Long Wen, Xinyu Li, and Liang Gao. A transfer convolutional neural network for fault diagnosis based on resnet-50. *Neural Computing and Applications*, 32(10):6111–6124, 2020. 7, 8

[64] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 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. 3

[65] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 2

[66] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *Advances in Neural Information Processing Systems*, 35:9125–9138, 2022. 3

[67] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 3

[68] Yan Zeng, Xinsong Zhang, Hang Li, Jiawei Wang, Jipeng Zhang, and Wangchunshu Zhou. X ˆ{2} 2-vlm: All-in-one pre-trained model for vision-language tasks. *IEEE transactions on pattern analysis and machine intelligence*, 46(5): 3156–3168, 2023. 5, 6, 7, 8

[69] Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Weidi Xie, and Yanfeng Wang. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*, 14(1):4542, 2023. 3

[70] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023. 5, 6, 7, 8

[71] Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. When radiology report generation meets knowledge graph. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12910–12917, 2020. 3

[72] Hong-Yu Zhou, Chenyu Lian, Liansheng Wang, and Yizhou Yu. Advancing radiograph representation learning with masked record modeling. *arXiv preprint arXiv:2301.13155*, 2023. 3

[73] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022. 3

[74] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 3

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

Supplementary Material

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.

Example (anonymized). We provide examples in the supplementary files, including imaging descriptions and diagnostic conclusions.

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 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 radiol-

Table 7. 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.

ogy. 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 7, 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 8. 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 8. 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		