Ultrasound-CLIP: Semantic-Aware Contrastive Pre-training for Ultrasound Image-Text Understanding

Supplementary Material

A. Ultrasonographic Diagnostic Attribute Framework (UDAF)

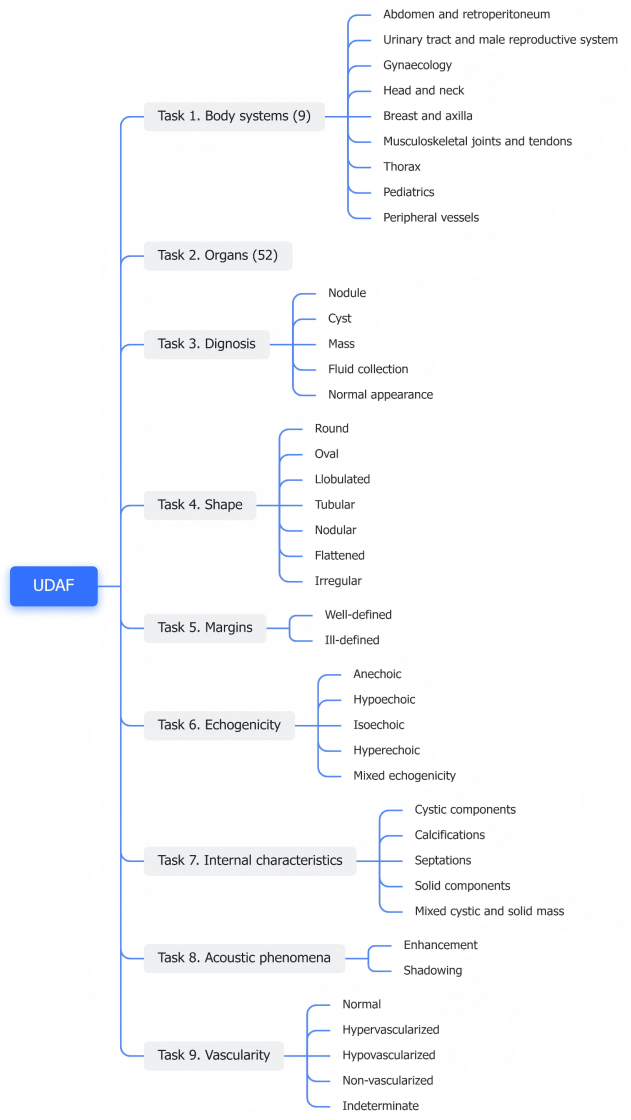


Figure 6. **Overview of the UDAF schema.** The framework decomposes ultrasound interpretation into nine standardized diagnostic dimensions. Note that Task 2 details are visualized in Figure 3 of the main paper.

The Ultrasonographic Diagnostic Attribute Framework (UDAF) serves as the semantic backbone of our dataset. As illustrated in Figure 6, UDAF provides a unified taxonomy that standardizes anatomical, morphological, and pathological attributes across heterogeneous ultrasound systems. By decomposing diagnostic-relevant information into hierarchical, multi-level attributes, UDAF enables consistent and granular annotation across diverse clinical cases.

B. Data Processing Pipeline

To ensure the reproducibility, transparency, and completeness of dataset construction, we provide a detailed description of our data construction pipeline, which consists of three sequential stages: (1) multi-source data collection, (2) data construction, and (3) image–text pairing and annotation. The complete workflow is illustrated in Figure 7, which highlights how heterogeneous online resources are transformed into a unified, clinically reliable ultrasound corpus.

Step 1: Multi-source Data Collection. To ensure comprehensive coverage and clinical diversity of ultrasound data, we systematically collect cases from five publicly accessible repositories, each providing distinct yet complementary characteristics and content types. All data collection adheres strictly to the terms of service and licensing agreements of the respective platforms, with all sources explicitly permitting academic and research use. The data sources are summarized as follows.

- *UltrasoundCases*[6]. This continuously maintained teaching-oriented platform developed by radiologists and sonographers provides a wide range of ultrasound imaging cases categorized by anatomical systems and organs. The website organizes data into 10 body systems and 60 organs, which served as the primary reference for constructing our HATU. The site exclusively contains ultrasound cases with rich visual and textual descriptions. In total, we collect 50,950 ultrasound images paired with corresponding textual annotations.
- *LITFL 100+ Ultrasound Quiz*[7]. The Life in the Fast Lane ultrasound quiz collection presents self-assessment materials comprising clinical scenarios, diagnostic questions, images, and key learning points. Although the number of ultrasound examples is limited, the accompanying text provides rich diagnostic reasoning and context. We select 111 representative cases from this source, including 213 videos and 102 still images, to enhance linguistic diversity and narrative quality.

[6] https://www.ultrasoundcases.info/
[7] https://litfl.com/top-100/ultrasound/

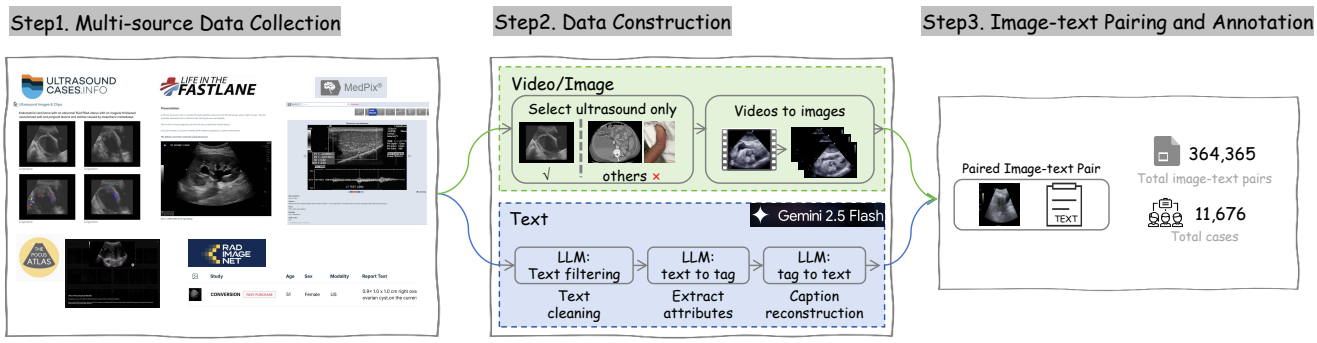


Figure 7. **The US-365K data construction pipeline.**

- *MedPix*[8]. The MedPix database, hosted by the U.S. National Library of Medicine, offers a large-scale, open-access collection organized by disease, anatomy, pathology, and imaging modality. We retrieve 40 ultrasound cases with 109 associated images and corresponding textual descriptions. These cases feature well-structured diagnostic summaries and cross-references, contributing to the medical accuracy of our corpus.
- *POCUS Atlas*[9]. This collaborative educational platform aggregates peer-reviewed point-of-care ultrasound (POCUS) cases submitted by clinicians worldwide. The website covers 13 organ categories and provides annotated imaging content with brief clinical notes. Due to limited textual descriptions, we primarily utilize 1,190 GIF sequences to enrich the dataset with dynamic ultrasound examples.
- *RadImageNet*[10]. RadImageNet is a GitHub-hosted open medical imaging resource. We incorporated the subset of publicly available ultrasound images for academic use. Although textual descriptions are sparse, the dataset provides standardized, high-quality images that help balance underrepresented organ classes. We ultimately include 3,000 image-text pairs from this source. All collected data are licensed under Creative Commons or equivalent open-access frameworks that explicitly permit academic research and educational use, enabling reproducible research within the scientific community.

Step 2: Data Construction. We first filter all non-ultrasound content to ensure strict domain relevance, retaining only ultrasound-specific videos and images. To integrate dynamic examinations, video clips are decomposed into static frames at 0.5-second intervals, balancing temporal diversity and redundancy. Although video samples account for a relatively small portion of the data, the extracted frames enrich the dataset with temporal cues such as probe motion, tissue deformation, and transient lesion appearances—capturing the dynamic nature of real-world sonographic examinations.

For textual refinement, we adopt a hybrid strategy combining large language model (LLM)–based automation and expert supervision. Specifically, we design structured prompts for Gemini 2.5 Flash to guide multi-stage text cleaning and reconstruction. The process begins with automatic filtering to remove irrelevant, redundant, or noisy descriptions, followed by medical attribute extraction to obtain structured entities such as anatomy, modality, and diagnostic findings. Subsequently, the extracted tags are reformulated into coherent and concise diagnostic captions through a "tag-to-text" generation step. All outputs are manually reviewed by trained annotators to ensure clinical accuracy, readability, and alignment with standard ultrasound reporting conventions.

Step 3: Image-text Pairing and Annotation. Following the standardized image and text refinement, each processed ultrasound image is explicitly aligned with its corresponding caption through a systematic pairing procedure. For multi-image cases or composite figures, we apply rule-based regular expression matching to detect subfigure indicators (e.g., "(A)", "(a)") and automatically associate each subfigure with its respective subcaption or narrative segment. This ensures fine-grained correspondence between localized visual regions and descriptive text, preserving diagnostic context across subcomponents.

Through this alignment pipeline, we establish 364,365 image–text pairs from 11,676 clinical cases, encompassing both static and video-derived ultrasound frames. The resulting dataset captures rich intra-organ variability and inter-system diversity, forming a coherent multimodal corpus suitable for large-scale pretraining and downstream clinical analysis.

C. Data Quality

To ensure the reliability and consistency of annotations across different anatomical systems, we implement a structured multi-stage quality control procedure aligned with

[8] https://medpix.nlm.nih.gov
[9] https://www.thepocusatlas.com/
[10] https://app.radimagenet.com/

Table 4. **Expert confidence ratings.** Self-reported confidence scores (1–5) from the three verifying medical experts across anatomical specialties. Higher values indicate stronger domain expertise.

Specialty	Expert 1	Expert 2	Expert 3
Liver	4	5	4
Kidney	4	4	5
Bladder	4	3	4
Scrotum	3	4	4
Uterus	4	4	3
Adnexa	3	4	4
Thyroid	5	4	4
Lymph nodes	4	4	4
Breast	4	3	4
Shoulder	3	4	3
Knee	3	4	4
Pleural space	4	4	4
Lung	4	4	4
Neonatal brain	3	3	4

the dataset construction workflow. During the annotation phase, three medical annotators examine outputs from the data curation pipeline, including the filtering of non-ultrasound content, the extraction of video frames at 0.5-second intervals, and the automatic UDAF-based label generation. Annotators verify the accuracy of extracted labels, the completeness of case information, and the correctness of rule-based subfigure matching used in image–text pairing. Cases with ambiguous anatomy, inconsistent textual records, or uncertain visual findings are flagged for further review.

In the review phase, three medical experts independently evaluate the refined image–text pairs. Each reviewer receives the ultrasound image, its caption, and the UDAF-aligned structured label set, and assesses two dimensions: *semantic alignment* (whether the caption accurately describes the acquired visual information) and *diagnostic consistency* (whether the labels match the clinical semantics expressed in the caption and image). To ensure broad anatomical coverage, samples from all nine ultrasound systems are included in the review pool. Furthermore, a dedicated confidence assessment is conducted in which experts rate their familiarity and interpretive confidence across 14 representative subspecialties, selected to provide balanced coverage while maintaining conciseness. The confidence scores, ranging from 1 to 5, are summarized in Table 4.

Pairs for which two or more experts identify issues in alignment or consistency are sent to a consensus adjudication stage. Annotators and experts collaboratively refine captions or labels until agreement is achieved. Across a randomly sampled subset of 5,000 image–text pairs from the dataset, the effective quality rate exceeds 93.2%.

D. US-365K Dataset Statistics

To provide a deeper understanding of the composition and linguistic characteristics of US-365K, we present extended statistical analyses in Figure 8. These visualizations complement the main paper and highlight the dataset's anatomical breadth, diagnostic diversity, and rich textual attributes.

D.1. Distribution of Diagnostic Findings

Figure 8(a) summarizes the frequency distribution of the top 15 diagnostic findings extracted from the refined captions. The histogram exhibits a characteristic long-tailed pattern:

- High-frequency findings (e.g., *fluid collection*, *mass*) provide abundant examples of common clinical presentations.
- Mid-frequency findings (e.g., *normal appearance*, *cystic components*, *increased vascularity*) represent a balanced mixture of organ- and lesion-level observations.
- Low-frequency but clinically relevant entities (e.g., *tubular/linear structures*, *irregular margins*) contribute rare but meaningful diagnostic cues.

This distribution confirms that the dataset captures both prevalent pathologies and infrequent but diagnostically important findings, forming a comprehensive representation of real-world sonographic examinations.

D.2. Caption Vocabulary Analysis

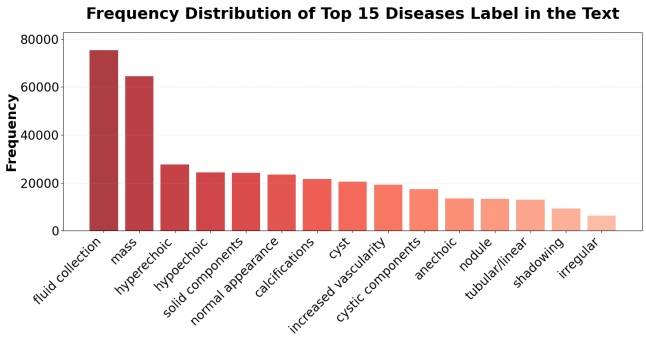
Figure 8(b) visualizes the linguistic space of the caption corpus through a word cloud generated from the normalized token distribution. The vocabulary demonstrates the following characteristics:

- A balanced mixture of general medical terminology (e.g., "ultrasound", "evaluation", "identified") and specialized sonographic descriptors (e.g., "hyperechoic", "cystic", "vascularity").
- High-frequency anatomical terms (e.g., "gallbladder", "retroperitoneal", "abdomen") consistent with the broad organ-level coverage defined by UHAT.
- Abundant descriptors of morphology and echogenicity (e.g., "anechoic", "shadowing", "components"), reflecting fine-grained imaging semantics.

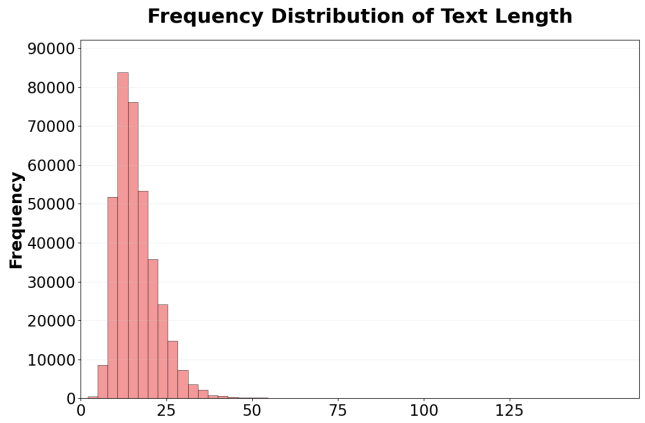
The vocabulary composition highlights both clinical specificity and linguistic diversity, supporting robust training for text-guided ultrasound understanding.

D.3. Caption Length Distribution

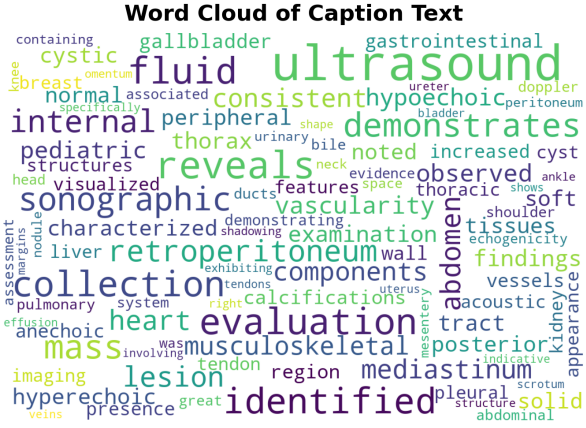
Figure 8(c) illustrates the caption length distribution across the dataset. The majority of captions fall within 10–25 words, consistent with the concise nature of radiology reporting. The distribution exhibits:



(a) Distribution of diagnostic findings.



(b) Word cloud of captions.



(c) Caption length distribution.

Figure 8. **Dataset statistics of the US-365K corpus.** (a) Frequency distribution of the most common diagnostic findings extracted from the curated text. (b) Word cloud summarizing the vocabulary usage in the full caption corpus. (c) Distribution of caption lengths, reflecting the conciseness and variability of clinical reporting.

- A mode around 15 words, corresponding to compact diagnostic summaries.
- A moderate tail toward longer captions (>40 words), typically arising from multi-lesion descriptions or more detailed clinical reasoning.
- Very few extremely short captions, due to our multi-stage reconstruction process designed to ensure descriptive completeness while avoiding redundancy.

This distribution demonstrates that US-365K preserves both the brevity and variability inherent to clinical ultrasound documentation.

D.4. Summary

Together, the visualizations in Figure 8 show that US-365K offers:

- Anatomical comprehensiveness: 9 body systems and 52 organs;
- Diagnostic diversity: coverage spanning common and rare findings;
- Linguistic richness: well-structured captions aligned with real-world clinical reporting.

These properties collectively make US-365K a strong foundation for ultrasound-focused vision–language pre-training, structured medical understanding, and downstream diagnostic reasoning tasks.

E. US-365K Dataset Splits

To ensure strict patient-level isolation, robust generalization assessment, and transparent reproducibility, we adopt a case-level split protocol and provide comprehensive corpus statistics. Unless otherwise specified, all analyses and evaluations in the main paper adhere to this protocol.

E.1. Case-level Split Protocol and Rationale

Ultrasound studies are inherently case-centric: multiple frames (including video-derived frames) within a single clinical case may share anatomy, pathology, scanning plane, and narrative context. Random image-wise splitting can therefore induce inadvertent leakage and optimistic bias. To prevent this, we perform splits at the case level, ensuring that no images or captions from the same clinical case appear across different splits (train/validation/test). This design supports faithful evaluation of generalization to unseen cases and mirrors realistic deployment scenarios.

We adopt a 6:2:2 split (training:validation:test) at the case level, stratified over the combined multi-source corpus to preserve source diversity in each partition. Table 5 summarizes the split sizes. The final materialized image–text pairs per split are reported in Table 6.

E.2. Stratification and Leakage Prevention

We construct splits by sampling at the case level with the following principles: (1) mutual exclusivity of cases across splits, (2) preservation of per-source composition to first

Table 5. Case-level split overview for US-365K. Splits are mutually exclusive by case.

Partition	#Cases	Proportion
Train	7,005	60.0%
Validation	2,336	20.0%
Test	2,335	20.0%
Total	11,676	100%

Table 6. Image–text pair counts per split for US-365K.

Partition	#Image–Text Pairs	Proportion
Train	218,402	59.94%
Validation	74,044	20.33%
Test	71,919	19.73%
Total	364,365	100%

order, and (3) invariance to within-case frame multiplicity (video-derived frames remain within the case they originate from). This prevents cross-split near-duplicates and leakage through subfigures, multi-frame snippets, or shared narratives.

To further reduce subtle leakage risks in composite figures (e.g., multi-panel images with shared captions), we associate all subfigures and subcaptions to the same parent case and enforce split-consistency at the parent level. The same policy applies to video-derived frames.

F. Prompts Design and Templates

We release the exact prompt texts used in all stages. To fit the two-column layout, we provide (1) short context bullets, (2) the full template in a compact code block, and (3) JSON schemas as inline lists. Placeholders in braces are programmatically injected.

F.1. Caption Reconstruction Prompts

This pipeline has three templates: Stage 1 (Anatomy tags), Stage 2 (Lesion attributes), and (3) Tag-to-Caption generation. All prompts enforce closed-world extraction, taxonomy-locked labels, and traceable rationale.

F.1.1. Stage 1: Anatomical Tag Extraction Prompts

Goal: Extract UHAT-aligned two-level anatomy (Body system, Organ).

Placeholders: $\{classification\_tree\}$, $\{output\_format\}$, $\{description\}$.

Output JSON schema: $Anatomy\_Body\_system\_level$, $Anatomy\_Organ\_level$, $Explanation : string$.

The complete prompt template is visualized in Figure 9.

```
PROMPT_TEMPLATE = """
You are a medical data annotator specializing in ultrasound imaging. Your task is to extract
structured anatomical location information from the input medical description.
Annotation Objective:
Extract two levels of anatomical location:

1. Anatomy_Body_system_level: The body system that the anatomical structure belongs
   to (first-level category in the classification tree).
2. Anatomy_Organ_level: The specific organ, region, or structure mentioned (second-level
   category in the classification tree).

Classification Tree (structured JSON):
{classification_tree}
Annotation Rules:

1. Only annotate locations explicitly mentioned or clearly implied in the input text. Do
   not infer or assume information.
2. If a body part belongs to a specific system (e.g., "finger" → "musculoskeletal system"),
   both levels must be annotated.
3. If the same structure belongs to multiple systems (e.g., "bone" → "musculoskeletal
   system", and also general involvement), assign both applicable values.
4. In ambiguous cases, leave the field empty.

Output Format:
{output_format}
Input:
{description}
"""
OUTPUT_FORMAT = """
{
"Anatomy_Body_system_level": ["..."],
"Anatomy_Organ_level": ["..."],
"Explanation": "..."
}
"""
```

Figure 9. Prompts for Stage 1: Anatomical Tag Extraction

F.1.2. Stage 2: Lesion Attribute Extraction Prompts

Goal: Extract UDAF-aligned attributes across remaining seven dimensions.

Placeholders: $\{attribute\}$, $\{output\_format\}$, $\{description\}$.

Rule: The model is instructed to extract only explicitly stated or clearly implied information. All labels must be from the closed sets.

The complete prompt template is visualized in Figure 10.

F.1.3. Stage 3: Tag-to-Caption Generation Prompts

Goal: Generate three clinically faithful sentences from tags, strictly adhering to the input features to prevent hallucination.

Placeholders: $\{Body\_system\_level\}$, $\{Organ\_level\}$, $\{diagnosis\}$, $\{shape\}$, $\{margin\}$, $\{echogenicity\}$, $\{internal\_features\}$, $\{posterior\_features\}$, $\{vascularity\}$, $\{Explanation\}$.

In scenarios where only anatomical tags (system/organ) are present, the model is instructed to generate brief, generic descriptive statements. The full prompt template is visualized in Figure 11.

```
PROMPT_TEMPLATE = """
You are a highly specialized medical data annotator with advanced expertise in ultrasound imaging analysis.
Your objective is to extract structured lesion attributes from the provided ultrasound report or description. The goal is to map the textual information into a standardized format while ensuring strict adherence to the defined rules and classifications.

1. Extraction Rules:
   Extract only the attributes explicitly stated or clearly implied in the input text. Avoid assumptions or inferences that are not grounded in the description provided.
   For each attribute, select the most appropriate label(s) from the fixed classification options. If multiple labels apply to a single attribute (e.g., "Shape" could be both "round" and "lobulated" if the text supports it), include all relevant labels in the output array for that attribute.
2. Handling Missing Values:
   If any attribute is missing, ambiguous, or not mentioned in the text, leave its corresponding value blank or an empty array in the JSON.
   Do not infer or fill in attributes based on prior medical knowledge or common imaging patterns. All extracted information must be directly supported by the input text.
3. Response Format:
   Your response must strictly adhere to the JSON structure specified in the "Output Format" section further below.
   General explanations, summaries, or additional commentary outside this defined JSON structure are not allowed. The "Explanation" field within the JSON has specific content requirements, as detailed in rule 5.
4. Content of the "Explanation" Field:
   In the "Explanation" field of the output JSON, you must explain for all extracted labels (across all attributes like Diagnosis, Shape, Margins, etc.):
   a. From which specific part(s) of the input text each label was derived.
   b. How each label was inferred or reasoned from that part of the text.
   This explanation should cover every label listed in the other fields of the JSON output, including instances where multiple labels are provided for a single attribute.

The attributes and their possible values are defined as follows:
{attributes}
Output Format:
{output_format}
Input:
{description}
"""
```

```
OUTPUT_FORMAT = """
{
"Diagnosis": ["..."],
"Shape": ["..."],
"Margins": ["..."],
"Echogenicity": ["..."],
"InternalCharacteristics": ["..."],
"PosteriorAcoustics": ["..."],
"Vascularity": ["..."],
"Explanation": "..."
}
"""
```

Figure 10. Prompts for Stage 2: Lesion Attribute Extraction

F.2. Multi-task Ultrasound Classification Prompts

We instantiate CLIP-style text prompts per task with concise, taxonomy-aligned descriptions. The mapping below defines the exact text strings used during evaluation. For space, we provide the full dictionary as a compact listing; each key is a class label, and the value is the prompt text.

- **"Task 1":** {
  "Abdomen and retroperitoneum": "a ultrasound image of Abdomen and retroperitoneum",
  "Urinary Tract and male reproductive system": "a ultrasound image of Urinary Tract and male reproductive system",

```
PROMPT_TEMPLATE = """
You are a professional medical AI assistant trained in radiology. Given a set of ultrasound findings, generate a clinically appropriate and natural-sounding medical report sentence describing the observed lesion. Ensure all medical terms are accurate and follow standard radiology reporting style.
Please use the following findings to generate your description:
Body_system_level: {Body_system_level} # e.g., "head and neck", "abdomen"
Organ_level: {Organ_level} # e.g., "thyroid", "kidney"
Diagnosis: {diagnosis} # e.g., "nodule", "cyst"
Shape: {shape} # e.g., "round", "irregular"
Margin: {margin} # e.g., "well-defined", "ill-defined"
Echogenicity: {echogenicity} # e.g., "hypoechoic", "mixed echogenicity"
Internal Features: {internal_features} # e.g., "septations", "calcifications"
Posterior Features: {posterior_features} # e.g., "enhancement", "shadowing"
Vascularity: {vascularity} # e.g., "increased vascularity", "no vascular signal"
Explanation:{Explanation}
Generate 3 diverse but medically accurate report sentences using this information. Each version should vary in tone, style, or focus, but all must retain correctness and reflect the findings, while ensuring strict adherence to the defined rules and classifications.
Additional Rules:

1. Strict Adherence to Provided Information: Generate captions only based on the information explicitly provided in the input fields. Do not infer, guess, or add any details not present in the input.
2. Conciseness with Limited Information: If only minimal information is provided (e.g., only the `Body_system_level` or `Organ_level` tags are filled), the generated captions should be correspondingly brief and general. For example, if only the organ is specified, a suitable caption might be "This is an ultrasound image of the [Organ_level]." or "Ultrasound evaluation of the [Organ_level]."
3. The output sentences should not contain the without detailing text, such as "without detailing any specific lesion".

Output Format(Example):
{{
"sentence1": "A well-defined, hypoechoic nodule with internal septations and posterior enhancement is noted. No vascular signal is detected on Doppler imaging.",
"sentence2": "Ultrasound reveals a localized, oval-shaped lesion showing mixed echogenicity and calcifications, accompanied by acoustic shadowing.",
"sentence3": "The lesion demonstrates irregular borders and heterogeneous echotexture, with internal solid and cystic components and increased vascularity."
}}
"""
```

Figure 11. Prompts for Stage 3: Tag-to-Caption Generation

  "Gynaecology": "a ultrasound image of Gynaecology",
  "Head and Neck": "a ultrasound image of Head and Neck",
  "Breast and Axilla": "a ultrasound image of Breast and Axilla",
  "Musculoskeletal Joints and Tendons": "a ultrasound image of Musculoskeletal Joints and Tendons",
  "Thorax": "a ultrasound image of Thorax",
  "Pediatrics": "a ultrasound image of Pediatrics",
  "Peripheral vessels": "a ultrasound image of Peripheral vessels" }
- **"Task 2":** {
  "Liver": "a ultrasound image of Liver",
  "Gallbladder and bile ducts": "a ultrasound image of Gallbladder and bile ducts",
  "Pancreas": "a ultrasound image of Pancreas",
  "Spleen": "a ultrasound image of Spleen",
  "Appendix": "a ultrasound image of Appendix",
  "Gastrointestinal tract": "a ultrasound image of Gastroin-

testinal tract",
"Peritoneum mesentery and omentum": "a ultrasound image of Peritoneum mesentery and omentum",
"Retroperitoneum and great vessels": "a ultrasound image of Retroperitoneum and great vessels",
"Adrenal glands": "a ultrasound image of Adrenal glands",
"Abdominal wall": "a ultrasound image of Abdominal wall",
"Kidney and ureter": "a ultrasound image of Kidney and ureter",
"Bladder": "a ultrasound image of Bladder",
"Scrotum": "a ultrasound image of Scrotum",
"Penis and perineum": "a ultrasound image of Penis and perineum",
"Uterus": "a ultrasound image of Uterus",
"Adnexa": "a ultrasound image of Adnexa",
"Vagina": "a ultrasound image of Vagina",
"Thyroid gland": "a ultrasound image of Thyroid gland",
"Parathyroid glands": "a ultrasound image of Parathyroid glands",
"Salivary glands": "a ultrasound image of Salivary glands",
"Lymph nodes": "a ultrasound image of Lymph nodes",
"Ocular": "a ultrasound image of Ocular",
"Ear": "a ultrasound image of Ear",
"Larynx": "a ultrasound image of Larynx",
"Breast": "a ultrasound image of Breast",
"Axilla": "a ultrasound image of Axilla",
"Shoulder": "a ultrasound image of Shoulder",
"Elbow": "a ultrasound image of Elbow",
"Wrist and carpus": "a ultrasound image of Wrist and carpus",
"Fingers": "a ultrasound image of Fingers",
"Hip groin and buttock": "a ultrasound image of Hip groin and buttock",
"Knee": "a ultrasound image of Knee",
"Ankle": "a ultrasound image of Ankle",
"Foot": "a ultrasound image of Foot",
"Peripheral nerves": "a ultrasound image of Peripheral nerves",
"Soft tissues": "a ultrasound image of Soft tissues",
"Skull": "a ultrasound image of Skull",
"Pulmonary": "a ultrasound image of Pulmonary",
"Pleural space": "a ultrasound image of Pleural space",
"Heart and mediastinum": "a ultrasound image of Heart and mediastinum",
"Thoracic wall": "a ultrasound image of Thoracic wall",
"Pediatric abdomen and retroperitoneum": "a ultrasound image of Pediatric abdomen and retroperitoneum",
"Pediatric urinary tract": "a ultrasound image of Pediatric urinary tract",
"Pediatric scrotum": "a ultrasound image of Pediatric scrotum",
"Pediatric gynaecological pathology and infant breast": "a ultrasound image of Pediatric gynaecological pathology and infant breast",
"Pediatric head and neck": "a ultrasound image of Pediatric head and neck",
"Neonatal brain and spine": "a ultrasound image of Neonatal brain and spine",
"Infant hip and knee": "a ultrasound image of Infant hip and knee",
"Pediatric thorax": "a ultrasound image of Pediatric thorax",
"Peripheral arteries": "a ultrasound image of Peripheral arteries",
"Peripheral veins": "a ultrasound image of Peripheral veins",
"Dialysis fistula": "a ultrasound image of Dialysis fistula" }

- **"Task 3":** {
"nodule": "a nodule in an ultrasound image",
"cyst": "a cyst in an ultrasound image",
"mass": "a mass in an ultrasound image",
"fluid collection": "a fluid collection in an ultrasound image",
"normal appearance": "normal appearance in an ultrasound image" }
- **"Task 4":** {
"round": "a round lesion in an ultrasound image",
"oval": "an oval lesion in an ultrasound image",
"lobulated": "a lobulated lesion in an ultrasound image",
"tubular/linear": "a tubular or linear lesion in an ultrasound image",
"nodular": "a nodular lesion in an ultrasound image",
"flattened": "a flattened lesion in an ultrasound image",
"irregular": "an irregular lesion in an ultrasound image" }
- **"Task 5":** {
"well-defined": "a lesion with well-defined margins in an ultrasound image",
"ill-defined/indistinct": "a lesion with ill-defined/indistinct margins in an ultrasound image" }
- **"Task 6":** {
"anechoic": "an anechoic lesion in an ultrasound image",
"hypoechoic": "a hypoechoic lesion in an ultrasound image",
"isoechoic": "an isoechoic lesion in an ultrasound image",
"hyperechoic": "a hyperechoic lesion in an ultrasound image",
"mixed echogenicity": "a lesion with mixed echogenicity in an ultrasound image" }
- **"Task 7":** {

"cystic components": "a lesion with cystic components in an ultrasound image",
"calcifications": "a lesion with calcifications in an ultrasound image",
"septations": "a lesion with septations in an ultrasound image",
"solid components": "a lesion with solid components in an ultrasound image",
"mixed cystic and solid mass": "a mixed cystic and solid mass in an ultrasound image" }

- **"Task 8":** {
"enhancement": "a lesion with posterior acoustic enhancement in an ultrasound image",
"shadowing": "a lesion with posterior acoustic shadowing in an ultrasound image" }
- **"Task 9":** {
"reduced/diminished vascularity": "a lesion with reduced or diminished vascularity in an ultrasound image",
"normal/regular vascularity": "a lesion with normal or regular vascularity in an ultrasound image",
"no vascularity": "a lesion with no vascularity in an ultrasound image",
"increased vascularity": "a lesion with increased vascularity in an ultrasound image",
"indeterminate/inhomogeneous vascularity": "a lesion with inhomogeneous or indeterminate vascularity in an ultrasound image" }

F.3. Downstream Classification Prompts

For each dataset, we use a single-sentence, class-conditioned template to form zero-shot labels. In the templates below, the placeholder {`class`} is dynamically replaced with the specific category name of the dataset.

- BUSBRA: "a breast ultrasound image showing {`class`} lesion"
- GIST514-DB: "an endoscopic image showing {`class`}"
- BreastMNIST: "a pathology image showing {`class`} breast lesion"
- Breast: "a breast ultrasound image showing {`class`}"

G. Experiment Results

G.1. Downstream Tasks

The results shown in Table 7 validate the impact of pre-training strategies on cross-dataset transferability. Overall, general-purpose CLIP models exhibit stable performance in the zero-shot setting, yet a noticeable semantic domain gap remains when applied to medical ultrasound. In contrast, most medical-specific CLIP variants achieve stronger discriminative power under limited-parameter learning and full fine-tuning, demonstrating the importance of medical text–image priors.

Among all compared methods, Ultrasound-CLIP consistently shows more balanced and robust generalization across datasets and evaluation protocols. In the zero-shot setting, it achieves the highest accuracy on GIST514-DB, reaching 53.89%, respectively, indicating that meaningful alignment between ultrasound structures and lesion semantics can be obtained without downstream training. In the LP setting, Ultrasound-CLIP attains the best or near-best performance on three datasets and achieves the top average ranking of 1.5, highlighting its suitability for low-annotation regimes. Under full fine-tuning, the model remains competitive, obtaining the highest accuracy on both BreastMNIST and Breast, and reaching the overall best average ranking.

Although several medical CLIP models outperform others in isolated tasks, Ultrasound-CLIP demonstrates more consistent cross-domain robustness across the four datasets and three evaluation settings. This stability stems from its large-scale pre-training corpus curated specifically for ultrasound understanding, enabling superior representation of anatomical structures, acoustic patterns, and lesion morphology compared to general vision–language models and existing medical multimodal models. Overall, the results indicate that Ultrasound-CLIP provides a reliable foundation model with strong transferability for a wide range of real-world ultrasound intelligence applications.

G.2. Efficiency Analysis

The results shown in Table 8. Ultrasound-CLIP (176.28M Params) is almost 5×smaller than SigLIP (877.96M), indicating that performance gains are not due to model scale. Notably, Ultrasound-CLIP achieves 74.97 FPS, surpassing all baselines, validating its feasibility for real-time deployment.

G.3. Robustness to Sparse Captions.

We group samples by caption sparsity (Table 9). The model maintains robust accuracy with different textual supervision, confirming its applicability to real-world clinical data.

G.4. Case Studies

Case Study: Probabilistic Alignment in Unstructured Diagnosis. Figure 12 illustrates a diagnostic case where the ground truth involves ambiguity between the labels "Mass" and "Fluid collection". Ultrasound-CLIP correctly identifies "Mass" as the primary label, aligning with the radiological description's emphasis on anatomical and shape attributes. However, what sets the model apart is its ability to capture the probabilistic proximity of alternative labels. Specifically, "Fluid collection," although not predicted as the top label, is ranked second, with nearly comparable confidence. This highlights the model's capacity for nuanced reasoning within diagnostic uncertainty.

Table 7. **Downstream task generalization comparison across 4 ultrasound datasets.** We report the accuracy performance under ZS, LP, and FT settings. * Models cannot be retrained due to unavailable code.

Method	BUS-BRA			GIST514-DB			BreastMNIST			Breast			Average Rank		
	ZS	LP	FT	ZS	LP	FT	ZS	LP	FT	ZS	LP	FT	ZS	LP	FT
General CLIP															
CLIP	51.49	70.16	86.15	49.03	62.58	76.13	28.33	73.92	90.38	50.77	69.66	84.19	6.25	6.5	3.75
SigLIP*	61.12	69.80	85.61	48.44	67.10	**81.40**	51.76	78.85	86.54	54.74	74.36	84.62	4.5	5	4
MetaCLIP	**65.97**	67.85	84.19	48.05	58.06	73.55	41.37	73.72	**91.67**	71.97	67.52	84.62	4.25	8	4.25
Medical CLIP															
PMC-CLIP	35.25	75.31	82.06	50.10	63.87	68.39	**69.87**	79.49	87.82	29.36	72.22	81.20	5	4.75	7.5
MedCLIP*	36.21	71.94	**87.21**	51.36	68.39	72.26	28.44	78.85	86.54	**74.49**	67.95	89.32	4.25	4.75	4.25
UniMed-CLIP	65.14	77.44	84.10	48.52	65.81	74.84	66.28	86.54	91.03	68.42	76.92	89.32	3.5	3	3.75
BiomedCLIP	33.33	77.44	84.19	51.56	**69.03**	74.84	47.69	84.62	89.10	52.44	**79.49**	88.89	5.25	1.75	4.25
Ultrasound-CLIP	54.93	**78.86**	84.55	**53.89**	68.39	72.90	49.10	**88.46**	**92.95**	70.26	77.35	**91.45**	**3**	**1.5**	**3**

Table 8. **Efficiency analysis.** We report the efficiency analysis results.

Model	CLIP	SigLIP	PMC CLIP	UniMed CLIP	Med CLIP	Meta CLIP	Biomed CLIP	Ultrasound CLIP
Params (M)	102	878	200	196	133	151	196	**176**
FLOPs (G)	7.36	326.7	13.5	33.0	4.73	4.89	25.0	26.9
FPS	44.4	16.8	36.1	45.7	33.9	32.6	43.5	**75.0**

Table 9. **Robustness to sparse captions.**

#Tasks	1	2	3	4	5	6	7	8	9
#Samples	10	25972	26094	13145	5872	1485	225	6	2
Acc (%)	80.00	45.72	59.40	54.30	45.61	50.25	48.19	37.50	61.11

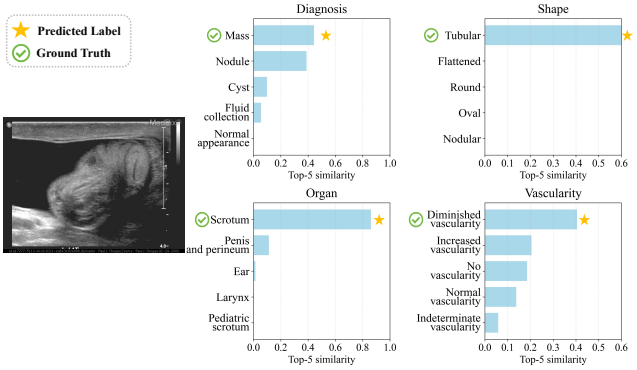


Figure 12. Probabilistic alignment in unstructured diagnosis.

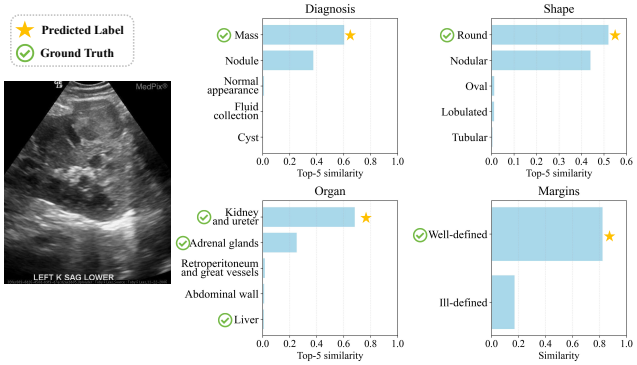


Figure 13. Multi-label clinical insight.

Additionally, supporting diagnostic attributes predicted by the model reinforce its alignment with the clinical context. For instance, "Rounded" for shape and "Well-defined" for margins are consistent with "Mass" but can also correspond to the fluid-filled presentation of certain cystic lesions. By effectively balancing the structured reasoning of primary and secondary outputs, Ultrasound-CLIP exhibits interpretability and robustness that mirror human diagnostic reasoning under conflicting or partially ambiguous cases. These results demonstrate the model's probabilistic reasoning depth and suggest its applicability across diverse clinical scenarios, even where ground truth spans multiple diagnostic possibilities.

Case Study: Multi-label Clinical Insight. In Figure 13, we present a case that showcases the adaptability of Ultrasound-CLIP to a complex diagnostic scenario involving multi-label clinical annotations. The ground truth includes multiple valid diagnoses, notably "Nodule" and "Cyst", reflecting the heterogeneous nature of this lesion's presentation. While our model is designed to predict a single primary diagnosis, its ranked predictions for supporting labels demonstrate high clinical alignment. Specifically, "Nodule" was identified as the primary label with a top confidence score, matching the main diagnosis in the report. Notably, "Cyst" while not the top-ranked prediction, is captured within the higher-ranked outputs. This demonstrates

that the model's probabilistic reasoning effectively maintains relevance across alternative diagnostic interpretations.

More critically, the supporting attributes predicted for "Nodule" (e.g., "Well-defined" for margins and "Hypoechoic" for echogenicity) are also highly consistent with the characteristics of "Cyst", indicating a nuanced understanding of overlapping diagnostic features. This case exemplifies Ultrasound-CLIP's ability not only to prioritize a single primary label but also to implicitly reflect the multi-label complexity inherent in clinical practice. The results further underscore the framework's capability to capture structured relationships between diagnostic attributes and provide clinically coherent reasoning, even in single-label prediction tasks.