Supplementary Material: Vision-Language Models Encode Clinical Guidelines for Concept-Based Medical Reasoning

Mohamed Harmanani[1,4], Bining Long[1*], Zhuoxin Guo[1,4*], Paul F.R. Wilson[1,4], Amirhossein Sabour[3], Minh Nguyen Nhat To[2], Gabor Fichtinger[1], Purang Abolmaesumi[2†], Parvin Mousavi[1,4†]
[1]Queen's University [2]University of British Columbia [3]McMaster University [4]Vector Institute
mohamed.harmanani@queensu.ca

A. Additional Experimental Details

A1. Code and configuration files

Our code and configuration files are publicly available at: github.com/mharmanani/medcbr.

A2. Datasets and Concepts

BUS-BRA [1] is a publicly available breast ultrasound dataset containing 1,875 images from 1,064 patients with biopsy-proven pathology labels and BI-RADS 2–5 assessments. Images were acquired retrospectively at the University of Brasília using a variety of scanners and linear-array transducers (5–14 MHz), resulting in heterogeneous imaging conditions characteristic of routine clinical practice. Lesions include masses abnormalities and are accompanied by radiologist-verified segmentation masks. All metadata were de-identified and curated to remove duplicates and low-quality scans. Dataset splits are constructed at the patient level to avoid data leakage. For this study, a trained radiologist annotated each lesion with 15 clinically relevant BI-RADS concepts to support concept-based modeling. The 15 concepts are shown in Table 1.

BrEaST [4] is a public breast ultrasound dataset containing 256 expertly annotated B-mode images with pixel-level lesion masks and BI-RADS descriptors. Images originate from multiple clinical settings and were curated to ensure consistent image quality, standardized acquisition views, and reliable expert annotations. In this work, BrEaST is used to augment BUS-BRA by providing additional training examples. Only images with complete annotations and verified pathology labels are included. Concept labels were provided by the authors of the dataset, and are the same as the ones highlighted in Table 1.

Concept Category	Concepts
Posterior features	(1) shadowing; (2) enhancement
Other findings	(3) halo; (4) skin thickening; (5) calcifications
Margins	(6) circumscribed; (7) indistinct; (8) angular; (9) microlobulated; (10) spiculated
Shape	(11) regular
Echogenicity	(12) hypoechoic; (13) hyperechoic; (14) heterogeneous; (15) cystic

Table 1. Clinically annotated BI-RADS concepts used for concept-based modeling.

CBIS-DDSM [5] is a curated subset of the Digital Database for Screening Mammography and provides ROI-level patches for benign and malignant lesions, each paired with pathology-confirmed labels. The dataset includes calcification and mass subsets with standardized training and test partitions. For each ROI, structured clinical concepts such as margin, shape, breast density, calcification shape, and calcification distribution are available. Only cropped ROIs are used in this work, and the mass and calcification subsets were merged to create a single unified dataset of 3500 images. As done in BUS-BRA, dataset splits for each fold were constructed at the patient level.

CUB-200-2011 [6] is a fine-grained natural image dataset containing 11,788 bird images spanning 200 species. Each image is annotated with 312 binary part- and appearance-based attributes covering shape, color, and texture. We adopt the standard train/test split provided by the dataset's creators and use a reduced concept bank of 112 concepts, as is commonly done in the literature [2, 3, 7]. This dataset serves as a natural image benchmark to evaluate the generality of MedCBR's concept grounding and rea-

*denotes equal contribution (with interchangeable order)
†denotes co-senior authorship

```
BIRADS_REPORTING_GUIDELINE_US = """
SUCCINCT DESCRIPTION OF THE OVERALL BREAST COMPOSITION (screening only)

Tissue composition patterns can be estimated more easily in the large FOVs of automated US scans but
    can also be discerned in the small FOV of a handheld US scan. The three US descriptors for tissue
    composition described earlier in the US lexicon, 'homogeneous
background echotexture-fat,' 'homogeneous background echotexture-fibroglandular,'
and 'heterogeneous background echotexture' (Table 3) (below) correspond loosely to
the four density descriptors of mammography and the four fibroglandular tissue descriptors of MRI. At
    US, breast tissue composition is determined by echogenicity. Subcutaneous fat, the tissue relative
    to which echogenicity is compared, is medium gray and darker
than fibroglandular tissue, which is light gray. Heterogeneous breasts show an admixture
of hypoechoic and more echogenic areas. Careful real-time scanning will help differentiate a small
    hypoechoic area of normal tissue from a mass.

Table 3. Breast Tissue
Tissue Composition
a. Homogeneous background echotexture-fat
b. Homogeneous background echotexture-fibroglandular
c. Heterogeneous background echotexture

CLEAR DESCRIPTION OF ANY IMPORTANT FINDINGS

The description of important findings should be made, in order of clinical relevance, using
lexicon terminology, and should include: [...]
"""
```

```
BIRADS_DIAGNOSTIC_GUIDELINE_US = """
# BI-RADS Ultrasound Diagnostic & Stratification Guideline

## Scope & Purpose
Standardizes description and assessment of breast findings on ultrasound (US).
Descriptors (shape, orientation, margins, echo pattern, posterior features, associated findings) are
    integrated to assign a BI-RADS category communicating malignancy likelihood.
The most suspicious ('dominant') feature determines the category.

---

## 1. Mass: Shape
- **Oval / Round / Gently Lobulated** - favors benignity (e.g., fibroadenoma).
- **Irregular** - suspicious; upgrade to BI-RADS 4-5 depending on context.

## 2. Mass: Orientation
- **Parallel (Wider-than-tall)** - benign-leaning.
- **Not Parallel (Taller-than-wide)** - malignant-leaning; indicates tissue plane invasion.

## 3. Mass: Margins
- **Circumscribed** - benign (e.g., fibroadenoma, cyst).
- **Microlobulated / Indistinct / Angular / Spiculated** - suspicious for invasion. [...]
"""
```

Figure 1. Example guideline snippets used to condition the LVLM (reporting) and the LRM (diagnostic).

soning beyond medical imaging.

B. Integrating Domain Guidelines

In this section, we provide further details on the construction of domain guidelines and their integration with large language and reasoning models. The full guideline text is available in our code.

B1. BI-RADS Clinical Guideline

For breast cancer detection, we constructed the following domain guidelines to support synthetic report generation and clinical reasoning:

1. `BIRADS_REPORTING_GUIDELINE_US`,
2. `BIRADS_REPORTING_GUIDELINE_MG`,
3. `BIRADS_DIAGNOSTIC_GUIDELINE_US`,

```
{
   "Pomarine Jaeger": """
   SPECIES ID: Pomarine Jaeger (Stercorarius pomarinus)
   OVERALL IMPRESSION: Largest jaeger; bulky with heavy chest and broad wings; breeding adults with
       twisted spoon-shaped tail projections.
   KEY IDENTIFICATION FEATURES: Head: Dark cap; pale nape; heavy bill. Chest: Whitish to buffy chest
       with dark sides. Back/Wings: Dark back; broad wings with pale bases to primaries. Tail: Two
       thick twisted tail spoons in breeding adults. Legs/Feet: Dark.
   BEHAVIOR AND POSTURE: Powerful, aggressive flier; pirates food from gulls/terns; hunts lemmings on
       tundra.
   HABITAT CONTEXT: High Arctic tundra for breeding; winters offshore worldwide.
   SIMILAR SPECIES: Parasitic Jaeger smaller; Long-tailed Jaeger slimmer with long streamers.
   DIAGNOSTIC FIELD MARKS: Large bulky build, twisted tail spoons in breeding, heavy flight. """,

   "Blue Jay": """SPECIES ID: Blue Jay (Cyanocitta cristata)
   OVERALL IMPRESSION: Bold, noisy corvid; blue above, white below, black necklace; common in eastern
       North America.
   KEY IDENTIFICATION FEATURES: Head: Blue crest; black collar; stout black bill. Chest: White chest
       and belly. Back/Wings: Blue back with black barring; white wing patches. Tail: Long blue tail
       with black bars and white tips. Legs/Feet: Dark.
   BEHAVIOR AND POSTURE: Noisy calls; mimics hawks; caches acorns; social at feeders.
   HABITAT CONTEXT: Woodlands, parks, suburbs in eastern North America.
   SIMILAR SPECIES: Florida Jay, Western Scrub-Jay; Blue Jay larger with crest.
   DIAGNOSTIC FIELD MARKS: Blue crest, black necklace, barred wings and tail, noisy calls. """, [...]
}
```

Figure 2. Example snippets of the field guide used to prompt the LVLM and LRM for CUB-200.

Concept Category	Concepts
Mass Shape	(1) regular (round/oval); (2) irregular; (3) lobulated
Mass Margins	(4) circumscribed; (5) ill-defined; (6) spiculated; (7) obscured; (8) microlobulated
Calcification	(9) pleomorphic; (10) amorphous; (11) fine linear; (12) branching; (13) vascular; (14) coarse; (15) punctate; (16) lucent-centered; (17) eggshell; (18) round; (19) regular; (20) dystrophic
Calcif. Distrib.	(21) clustered; (22) segmental; (23) linear; (24) scattered; (25) regional
Breast Density	(26) low density (27) moderate density (28) high density
Other Findings	(29) architectural distortion; (30) asymmetry; (31) lymph node

Table 2. Ground-truth concept labels for CBIS-DDSM.

4. `BIRADS_DIAGNOSTIC_GUIDELINE_MG`,

The two *reporting* guidelines describe how to structure a BI-RADS-compliant clinical report from ultrasound and mammography images respectively, including phrasing, ordering of findings, and required components (lesion description, assessment, and recommendation). These are used to steer the LVLM during text generation.

The two *diagnostic* guidelines specify the clinical implications of relevant BI-RADS concepts (e.g., which findings suggest benign or malignant pathology, how specific descriptors alter risk). These are used by the LRM to produce guideline-aware diagnostic narratives.

B2. Sibley-Inspired Field Guide

For the CUB-200 images, we provide a compact field-guide entry inspired by the Sibley bird identification manual. Each entry summarizes key visual traits (color, shape, pattern, behaviors) and their taxonomic relevance. These descriptions serve as lightweight analogues of clinical guidelines for non-medical domains. An example of the guideline can be seen in Figure 2.

B3. Prompting the LVLM

Prompt Structure. For BUS-BRA and CBIS-DDSM, the LVLM is conditioned using the appropriate reporting guideline, which provides a structured template for BI-RADS–compliant report generation. For CUB-200, the LVLM instead receives the corresponding Sibley-inspired field-guide entry for the species, offering a compact attribute-oriented template.

Prompting Strategy. We explored several prompting strategies to ensure that LVLM outputs were reliable and

```
concept_data = "Finally, you are given the
    following 'concepts' " \
           "that are present in the image.\n"

for i in range(len(selected_concepts)):
   if metadata["concepts"][i] == 1:
      name = named_concepts[i].replace("_", "
          ").capitalize()
      concept_data += f"{name}: 1\n"

prompt = f"""
You are given the following {modality} <image>.
    {auxiliary_data}
You are also given the following
    {type_of_guideline} guideline:

{GUIDELINE}

{concept_data}

Write a report based on the image, the guideline
    provided, and the concepts present in the
    image.
"""

messages = [{
   "role": "user",
   "content": [
      {"type": "image", "image": image},
      {"type": "text", "text": prompt},
   ],
}]
```

Figure 3. LVLM prompting strategy used in our experiments. The LVLM receives the image, the appropriate reporting guideline (BI-RADS or field-guide), the predicted concepts, and a final instruction describing the reporting task.

```
introduction = "You are given the final
    diagnostic prediction of an AI system, which
    is {diagnosis}. The system also detected the
    following concepts:\n"

concept_data = ""
for i, name in enumerate(named_concepts):
   score = metadata["concepts"][i] * 100
   if metadata["concepts"][i] >= 0.5:
      cname = name.replace("_"," ").capitalize()
      concept_data += f"{cname} ({score:.1f}%
          confidence)\n"
   else:
      if dataset_name == "BREAST_US" and name ==
          "regular_shape":
         concept_data += f"Irregular shape
             ({score:.1f}% confidence)\n"

# for CUB, this would say "field guide"
instructions = "Assuming the diagnosis is
    correct, explain the implications of these
    concepts according to the BI-RADS clinical
    guideline provided. Interpret each concept,
    assess agreement with the predicted
    diagnosis, infer the most likely BI-RADS
    category, and provide a recommended
    follow-up.\n"

reasoning_prompt = f"""{introduction}
{concept_data}
{instructions}
{GUIDELINE}
"""
```

Figure 4. LRM prompting strategy. The LRM receives the model's final prediction, the predicted concepts, a domain-specific diagnostic or field-guide guideline, and instructions describing how to construct a reasoned explanation grounded in those concepts.

grounded. First, we supplied only the image and requested a clinical report; this produced text that was stylistically coherent but frequently inaccurate or logically inconsistent. Incorporating the BI-RADS reporting guideline improved the structure and style of the output but did not fully correct factual errors. We therefore augmented the prompt with the ground-truth concept labels to anchor the LVLM to verifiable findings, as shown in Figure 3. For each configuration, outputs were evaluated by comparing the main assertions in the generated report with the clinical ground truth, and all reports were subsequently reviewed by a trained radiologist for fact-checking.

B4. Prompting the LRM

The LRM receives the predicted concepts from the concept-based model together with the appropriate *diagnostic* guideline. These guidelines specify the clinical or taxonomic implications of each concept (e.g., "spiculated margins increase suspicion for malignancy" or "a blue crest is characteristic of a blue jay"), enabling the LRM to generate coherent reasoning narratives that reflect domain conventions. By grounding its explanation in both the structured predictions and the diagnostic rules, the LRM produces interpretable statements aligned with established clinical or field-guide knowledge. A snippet of our LRM prompting strategy can be seen in Figure 4.

C. Hyperparameter Tuning

Original CBM. For the baseline concept bottleneck model, we performed grid searches over batch size, learning rate, data augmentations, loss formulations, and optimizers. We evaluated two loss functions: (i) $\mu\mathcal{L}_{\text{CE}}(y,\hat{y}) + \nu\mathcal{L}_{\text{CE}}(c,\hat{c})$ and (ii) $\mathcal{L}_{\text{CE}}(y,\hat{y}) + \mathcal{L}_{\text{MSE}}(c,\hat{c})$. Formulation (i) performed consistently better, and we adopted $\mu = 1.0$, $\nu = 0.8$. Across optimizers (SGD, Adam, AdamW), AdamW yielded the most stable results. All models were trained for 150 epochs with early stopping based on validation loss.

CLIP CBM. We evaluated several CLIP encoders (ViT-B/32, ViT-L/14, and RN50) as the visual backbone. CLIP

Metric	Description and Scoring Criteria
Concept Interpretation Score (CIntS)	**Definition:** Measures whether each predicted concept is interpreted correctly according to BI-RADS semantics. **Score:** $$\text{CIntS} = \frac{\text{\# correctly interpreted concepts}}{\text{\# predicted concepts}}$$
Concept Integration Score (CIgS)	**Definition:** Evaluates whether multiple concepts are integrated coherently and consistently with BI-RADS guidelines. **Scoring Levels:** **1.0** **Fully Correct:** All concept combinations follow BI-RADS guidelines; no contradictions. **0.75** **Mostly Correct:** Most interpretations are reasonable, but some combinations deviate from BI-RADS (e.g., many malignant cues and a few misinterpreted concept combinations, yet the overall impression is still malignant). **0.25** **Partially Correct:** Only a small portion is reasonable; several contradictions or guideline violations (e.g., several malignant cues and a few benign concept combinations, with a benign overall impression). **0.0** **Incorrect:** Concept combinations or guideline references are entirely inconsistent or unreasonable.
BI-RADS Assignment Score (BAS)	**Definition:** Measures whether the final BI-RADS category aligns with guideline-based decision criteria for the given case. **Scoring Levels:** **1.0** **Correct Assignment:** Correct BI-RADS category chosen based on BI-RADS decision logic. **0.8** **Near-Correct:** Unreasonable category, but correct malignant/benign implication (e.g., 4C vs. 5). **0.0** **Incorrect:** Category contradicts BI-RADS criteria and results in an incorrect implication.

Table 3. Rubric used by the radiologist for evaluating the clinical validity of model reasoning. CIntS measures accuracy of individual concept interpretation, CIgS measures the coherence of multi-concept integration, and BAS evaluates correctness of the final BI-RADS assignment.

ViT-L/14 consistently achieved the strongest performance and was therefore used as the default. We searched over the concept and label loss weights and retained $\mu = 1.0$ and $\nu = 1.0$. All other hyperparameters followed the original CBM setup.

Label-Free CBM. For LF-CBM, the primary hyperparameter is the strategy used to generate concept labels via GPT. Directly prompting GPT often resulted in redundant or noisy concepts, inflating the concept bank and reducing performance. We also tested bypassing GPT entirely and inserting curated concept labels before training using the authors' code. Performance varied by dataset, and we report the best-performing configuration in each case.

We additionally tuned the visual backbone and compared CLIP RN50 and CLIP ViT-L/14. The optimal backbone varied by dataset; we selected the best-performing one for each experiment.

AdaCBM. As with LF-CBM, the main hyperparameter in AdaCBM is the source of concept labels (GPT-generated vs. curated). On CUB-200, models using the GPT-derived concepts provided in the LF-CBM repository significantly outperformed those using the official ground-truth concept bank. We also tuned the visual backbone, comparing CLIP ViT-B/32 and CLIP ViT-L/14. For each dataset, we retained the backbone that yielded the highest validation performance.

AdaCBM uses a two-stage concept selection procedure: a t-test to filter non-informative concepts, followed by correlation-based redundancy removal. This process was often overly aggressive for CUB-200, removing nearly all concepts. We therefore tuned the *interpretability cutoff* associated with the t-test threshold to reduce over-pruning. When the available concept bank was small, we adopted a very permissive (low) cutoff to preserve a sufficient number

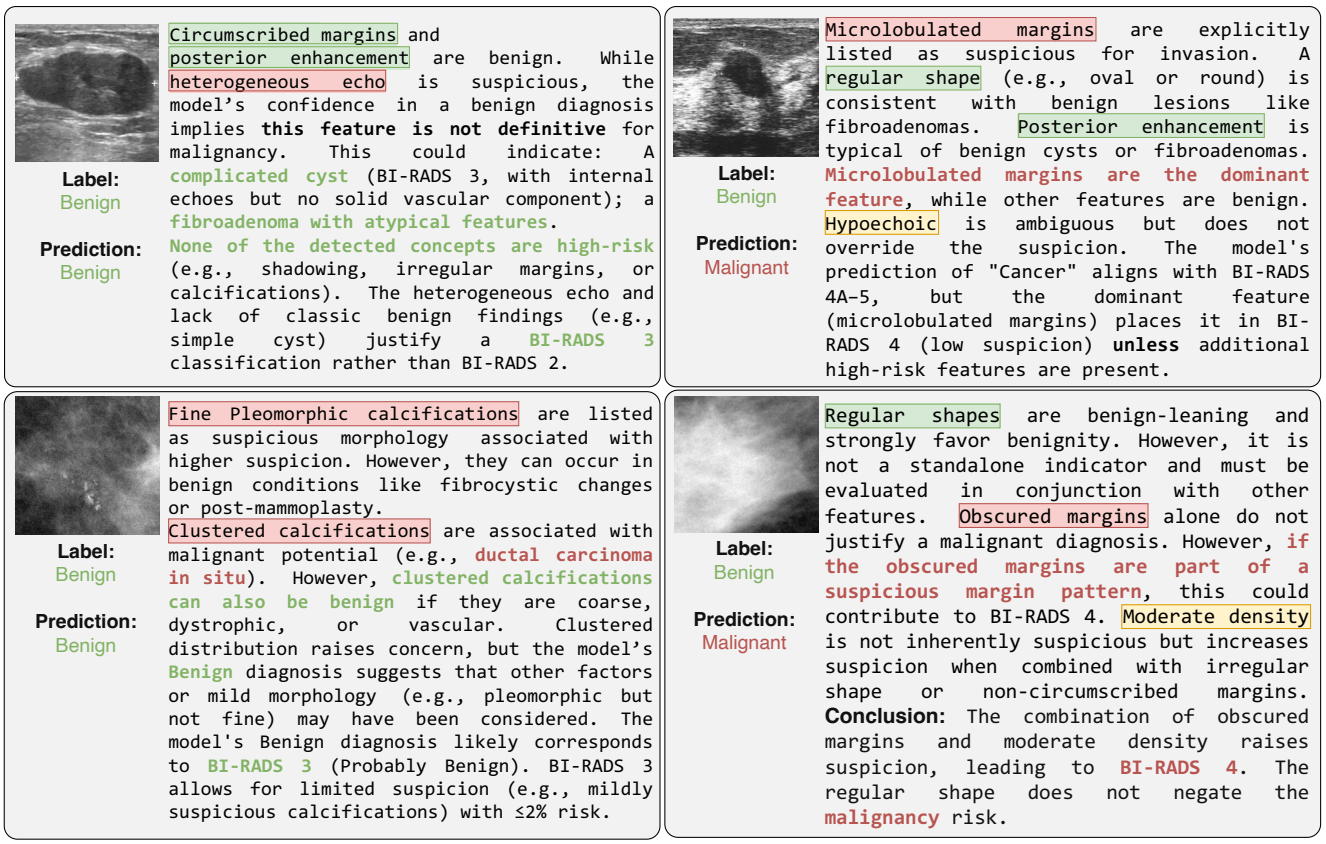

Figure 5. Additional examples of model reasoning on four challenging cases from BUS-BRA (top) and CBIS-DDSM (bottom). Malignant and benign-leaning clinical concepts are highlighted in red and green, and their corresponding reasoning steps are emphasized. Neutral concepts are highlighted in yellow.

of concepts and improve downstream accuracy.

For each model and dataset, we report results from the best-performing configuration.

D. Evaluation of Model Reasoning

D1. Radiologist Evaluation

To complement our quantitative reasoning metrics, we conducted an expert evaluation with an attending radiologist with specialized breast imaging training. The radiologist assessed 20 randomly selected cases drawn independently by a separate investigator to avoid bias. For each case, the radiologist reviewed the predicted concepts, the LRM-generated reasoning statement, and the final diagnostic assignment, while remaining blinded to the ground-truth pathology and overall dataset composition.

The assessment followed a structured evaluation rubric, which is outlined in full in Table 3. The rubric evaluates (i) the *Concept Interpretation Score (CIntS)*, measuring whether each concept is interpreted correctly within BI-RADS semantics; (ii) the *Concept Integration Score (CIgS)*, which evaluates whether multiple concepts are combined logically and conflicts between benign and malignant cues are resolved; and (iii) the *BI-RADS Assignment Score (BAS)*, which assesses whether the final diagnostic category is justified according to BI-RADS guideline criteria.

D2. Additional Examples of Model Reasoning

We include additional qualitative reasoning examples on complex cases from BUS-BRA and CBIS-DDSM (shown in Figure 5), as well as CUB-200 (shown in Figure 6) to illustrate how MedCBR integrates concept predictions with domain-specific guidelines. Each example includes the image, instances of predicted concepts, the LRM-generated reasoning statement, and the final classification. For each of BUS-BRA and CBIS-DDSM, we highlight both a success case (left) and a failure case (right). For CUB-200, we highlight 4 complex cases where the concept model predictions are either partially correct or entirely incorrect, and show the LRM weighing conflicting evidence to support a coherent conclusion.

Reasoning on medical images. In the top left image of Figure 5, the model predicts a benign outcome and the LRM correctly explains that an isolated suspicious feature

Label: Gray Kingbird

Prediction: Gray Kingbird

The model detected "Has bill shape hooked" (74%). The field guide mentions a "heavy bill," which might be hooked, so that's a match. Next, the wings. The model has several entries: brown (53%), iridescent (81%), black (89%). Darker wings could be black, but the model also has iridescent. The guide doesn't mention iridescent. The guide says "white to pale gray belly," which is smooth, not spotted. Wait, that's a conflict. Breast color: grey (56%). The guide says "white to pale gray belly," so grey is a possible match. **Conclusion:** The model's detection of heavy bill, dark eye mask, white underparts, and black wings supports the classification of Gray Kingbird. While some features conflict with the field guide, this may stem from image-specific variability, or misclassification.

Label: Florida Jay

Prediction: Florida Jay

The model detected "wing color green". However, the field guide says blue wings, that's a problem. Let me check other features. The model has "upper tail color white" with 89% confidence. The field guide says "blue tail." So the implications would be that the model's concepts are not accurate, but the prediction is correct, which could be due to other features being accurate. For example, the high confidence in "capped" likely refers to the blue head (a key diagnostic feature), which matches the field guide.

Conclusion: The model's prediction of Florida Scrub-Jay is justified by key diagnostic features, but conflicting concepts (e.g., green wings, white tail, spotted breast) suggest potential misinterpretations of color or texture.

Label: Orange Crowned Warbler

Prediction: Orange Crowned Warbler

The field guide says the back and wings are olive-green. However, the model has high confidence in red [wings]. But the model's underparts color olive is 84.84%, which might be the chest. The field guide says "yellow-olive chest," so maybe the underparts color olive is correct. The field guide mentions a faint orange crown, which might be detected as orange in the upper tail. So the model's detection of orange could align with the "faint orange crown" mentioned in the guide.

Conclusion: The model's predictions, while containing some features that diverge from the field guide, are consistent with the Orange-crowned Warbler when considering: The faint orange crown (interpreted as upper tail orange), Olive-green tones in upperparts and underparts.

Label: Winter Wren

Prediction: House Wren

The model detected "Has bill shape hooked" with high confidence. But the field guide says the House Wren has a "fine thin bill." Wait, hooked vs. thin? That seems conflicting. Next, wing colors. The model has white [wings]. The field guide mentions brown wings with fine barring. Underparts: The model has red [underparts] with 88% confidence. The field guide says "pale gray-brown with faint barring." That's a big discrepancy. The model says "notched tail". The field guide mentions "short, barred; often cocked upward."

Conclusion: The model's high confidence in hooked bill, red underparts, and notched tail strongly conflict with the House Wren's description. These may indicate misclassification, overfitting, or confusion with similar species.

Figure 6. Additional examples of model reasoning on 4 complex cases from CUB-200. Observations highlighted in orange contradict the guideline, whereas ones in blue are consistent. Key reasoning steps are emphasized in **blue** when they support the final conclusion and **red** when they do not.

can still appear in benign lesions. In the top right image, the model incorrectly predicts malignancy; although the final diagnosis is benign, this case presents mixed or misleading features, leading to a higher risk score. In the bottom left image, despite multiple suspicious concepts, the model predicts benign correctly, and the LRM explains how these findings may still be compatible with a low-risk assessment. In the bottom right image, the model incorrectly predicts malignancy, and some detected concepts contradict the visual appearance (e.g., regular shape), leading the LRM to assign an overly high BI-RADS category.

Reasoning on natural images. In Figure 6's top left example, the model detects the correct concepts, and generates a reliable explanation. In the subsequent two examples (top right and bottom left), the final prediction is correct, but the concepts contain multiple mistakes. These mistakes are then correctly identified by the LRM and highlighted to the user. Finally, in the bottom right case, the LRM rejects the concept model's prediction due to too many predicted concepts conflicting with the field guide.

E. Ethical Considerations

First, there is a risk that clinicians may place undue trust in model-generated narratives, even when predictions are incorrect. Our framework is intended to support expert judgment, not replace it. Second, LLMs are computationally expensive to train and deploy, raising concerns about their environmental impact. Future work may explore efficient alternatives such as small language models (SLMs) or distillation-based techniques. Finally, while our method greatly reduces the likelihood of hallucinated outputs by conditioning on auxiliary information (e.g., predicted concepts), the risk of factual inaccuracy remains, especially in cases where the concept model fails to output reliable predictions. Additional safeguards and validation strategies are

needed to ensure the reliability of generated narratives in clinical settings.

References

[1] Wilfrido Gómez-Flores, Maria Julia Gregorio-Calas, and Wagner Coelho de Albuquerque Pereira. Bus-bra: a breast ultrasound dataset for assessing computer-aided diagnosis systems. *Medical Physics*, 51(4):3110–3123, 2024. 1

[2] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020. 1

[3] Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. *Internation Conference on Learning Representations*, 2023. 1

[4] Anna Pawłowska, Anna Ćwierz-Pieńkowska, Agnieszka Domalik, Dominika Jaguś, Piotr Kasprzak, Rafał Matkowski, Łukasz Fura, Andrzej Nowicki, and Norbert Żołek. Curated benchmark dataset for ultrasound based breast lesion analysis. *Scientific Data*, 11(1):148, 2024. 1

[5] Rebecca Sawyer-Lee, Francisco Gimenez, Assaf Hoogi, and Daniel Rubin. Curated breast imaging subset of digital database for screening mammography (cbis-ddsm). *TCIA: The cancer imaging archive*, 2016. 1

[6] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. Caltech-ucsd birds-200-2011 (cub-200-2011). Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 1

[7] Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. *Internation Conference on Learning Representations*, 2023. 1