Knowledge-Driven Hypothesis Generation for Burn Diagnosis from Ultrasound with Vision-Language Model

Md Masudur Rahman[1][0000−0002−3633−0621]⋆, Mohamed El Masry[2,3], Gayle Gordillo[2,4], and Juan P Wachs[1]

[1] Edwardson School of Industrial Engineering, Purdue University, West Lafayette, IN 47907, USA
{rahman64,jpwachs}@purdue.edu

[2] McGowan Institute for Regenerative Medicine (MIRM), Pittsburgh, PA 15219, USA

[3] Department of Surgery, University of Pittsburgh School of Medicine, Pittsburgh, PA 15213, USA

[4] Department of Plastic Surgery, University of Pittsburgh School of Medicine, Pittsburgh, PA 15261, USA moelmasry@pitt.edu,gordillogm@upmc.edu

Abstract. Although vision-language models (VLMs) have achieved strong results in general computer vision tasks, their effectiveness in medical imaging remains limited—primarily due to their insufficient reasoning capabilities. In this work, we introduce KODER, a novel knowledge-driven reasoning framework aimed at improving diagnostic accuracy for ultrasound-based burn assessment. KODER integrates pre-trained VLMs with first-order logic (FOL) reasoning to generate interpretable diagnostic hypotheses. By combining rich experimental descriptions and clinical insights into a unified prompt, the framework produces multiple diagnostic hypotheses and refines them through iterative consistency checks using an SMT solver. The validated hypotheses are then used to support both surgical decision-making and detailed burn depth classification. We evaluate our approach on a retrospective dataset collected from a U.S. burn center, where it achieves significant performance gains—reaching up to 93% accuracy in surgical classification and 87% in fine-grained burn depth prediction. Additionally, incorporating techniques such as chain-of-thought reasoning, self-consistency, and explicit explanation generation further boosts both interpretability and diagnostic reliability. Our experiments span multiple state-of-the-art VLMs, including GPT-4o, GPT-4 Turbo, and Gemini 1.5 and Gemini 2.0, confirming the generalizability of KODER across architectures.

Keywords: Vision-Language Models · Ultrasound Imaging · Burn Diagnosis · Hypothesis Generation · Chain-of-Thought Reasoning · Self-Consistency · First-Order Logic

⋆ Corresponding author(s): rahman64@purdue.edu, jpwachs@purdue.edu

1 Introduction

Vision-language models (VLMs) have achieved significant success by integrating visual processing with natural language understanding, particularly in tasks where explicit reasoning is not a core requirement [1,17,10,9,13,12,7,22,8,5,21,14]. However, applying these models to complex medical scenarios—such as diagnosing burn severity from ultrasound scans—presents unique challenges. This difficulty arises from the fact that interpreting multimodal ultrasound data often involves a structured reasoning process rather than straightforward perception.

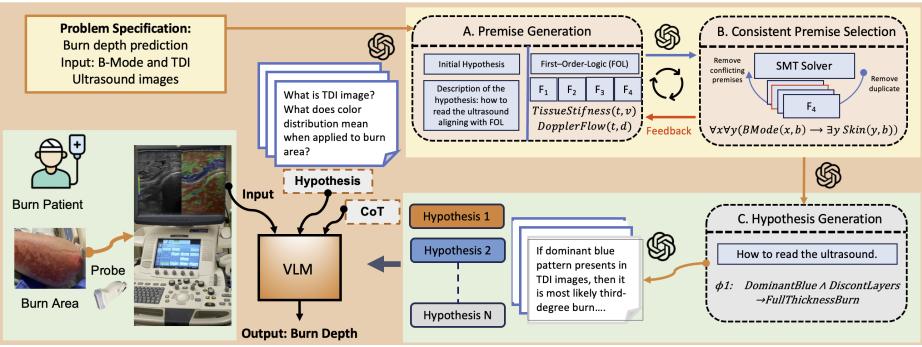

Fig. 1. Overview of KODER: **Kno**wledge-**D**riv**e**n **R**easoning Framework.

These challenges become even more pronounced when working with novel imaging modalities and small, domain-specific datasets. VLMs are typically pretrained on massive, diverse datasets [13], but such data is rarely available in specialized clinical settings due to privacy concerns, proprietary limitations, or lack of standardization. This issue is particularly relevant in emerging areas like ultrasound-based burn care, where imaging protocols are still evolving and not widely adopted in practice [18]. As a result, training large, dedicated models for such tasks is often infeasible. To overcome these limitations, we propose a new framework that adapts existing, *general-purpose* VLMs to medical imaging tasks by incorporating structured reasoning.

Another critical limitation of large language models (LLMs) is their tendency to generate explanations that may be ambiguous, inconsistent, or lack clinical clarity [20,19,3,4]. This stems from the probabilistic nature of LLM outputs and can be problematic in high-stakes applications like medical diagnosis, where interpretability and reliability are essential [16,2]. For instance, ultrasound modalities such as Tissue Doppler Imaging (TDI) [6] and B-mode imaging capture nuanced features—including color-coded motion patterns and tissue structures—that require both image understanding and expert-level reasoning to interpret correctly.

To address this, we introduce **KODER** (**Kno**wledge-**D**riven **R**easoning), a framework that combines the generative capabilities of LLMs with formal logical validation. As shown in Figure 1, our method begins with a rich textual

description of the imaging task, including details about modalities, scanning conditions, and patient context. The LLM uses this information to generate a diagnostic hypothesis along with a set of first-order logic (FOL) premises that encode the clinical rationale. These premises are then evaluated using an SMT (Satisfiability Modulo Theories) solver such as Z3 [11] to detect contradictions or ambiguities. This verification process enables iterative refinement of the hypothesis and logic until a consistent and clinically sound conclusion is reached.

We evaluate our framework on two downstream tasks: (i) binary classification to determine whether surgical intervention is needed, and (ii) a fine-grained, three-class burn depth prediction. Our experimental results on ultrasound datasets for burn diagnosis show that incorporating logically validated hypotheses leads to improved diagnostic accuracy. Across multiple state-of-the-art VLMs—including `GPT-4o`, `GPT-4 Turbo`, and `Gemini 1.5` and `Gemini 2.0` — KODER consistently improves performance. For instance, GPT-4o combined with KODER achieves up to **93% accuracy** in surgical decision-making and **87%** in burn depth classification. Moreover, incorporating chain-of-thought reasoning and self-consistency further boosts both accuracy and interpretability, highlighting the effectiveness of structured, knowledge-guided diagnostic modeling.

In summary, our main contributions are:

- We introduce a novel framework that integrates vision-language models with formal logic to produce interpretable diagnostic hypotheses.
- We address the limitations of standard LLM/VLM outputs by validating generated content through logical consistency checks.
- We demonstrate the clinical utility of our approach through improved performance on real-world diagnostic tasks in burn care.

2 Methodology

2.1 KODER Framework

This section presents the **K**n**o**wledge-**D**riv**e**n **R**easoning (KODER) framework, a method developed to formulate diagnostic hypotheses for predicting burn depth based on ultrasound imaging.

Problem Formulation and Input Description. The primary goal is to generate a global, dataset-level hypothesis for predicting burn depth based on ultrasound imaging. Instead of relying on raw image data, the method uses a detailed textual description that outlines both the imaging modalities and clinical rationale.

Let $\mathcal{T}$ denote the space of textual descriptions. We define $\mathcal{D}_{\text{exp}} \in \mathcal{T}$ as the *experimental setup description*, and $\mathcal{D}_{\text{clin}} \in \mathcal{T}$ as the *clinical context information*. For example, $\mathcal{D}_{\text{exp}}$ could be: "We employ Tissue Doppler Imaging (TDI) and B-mode ultrasound to predict burn depth. TDI provides color-coded velocity information, while B-mode offers structural imaging." Similarly, $\mathcal{D}_{\text{clin}}$ could be:

"Clinical protocols indicate that dominant blue patterns in TDI images and discontinuous layers in B-mode images are correlated with full-thickness burns."

We define the *PromptBuilder* function to merge these two sources of information into a single prompt p:

$$p = \text{PromptBuilder}(\mathcal{D}_{\text{exp}}, \mathcal{D}_{\text{clin}}) = \mathcal{D}_{\text{exp}} \oplus \mathcal{D}_{\text{clin}},$$

where $\oplus$ represents concatenation, and PromptBuilder $\in \mathcal{T}$. This prompt provides the LLM with sufficient contextual knowledge for hypothesis generation.

Hypothesis and Premise Generation. Using the prompt p, the language model M_θ (parameterized by θ) generates both a natural language hypothesis h and a corresponding set of first-order logic (FOL) premises Φ. To promote diversity in the outputs, sampling parameters such as temperature τ and top-p nucleus sampling p_{top} are varied. Formally:

$$(h, \Phi) = M_\theta(p \mid \tau, p_{\text{top}}),$$

where h is a natural-language hypothesis, e.g., "If a dominant blue pattern is observed in TDI images and B-mode images show discontinuous layers, then the burn is likely full-thickness." Additionally, $\Phi = \{\phi_1, \phi_2, \ldots, \phi_K\}$ is a set of FOL statements encoding clinical rules. For instance:

$$\phi_1 : \quad (\text{DominantBlue} \wedge \text{DiscontLayers}) \rightarrow \text{FullThicknessBurn}.$$

Consistency Verification via SMT Solver. To verify the internal consistency of the logical premises Φ, a satisfiability modulo theories (SMT) solver (e.g., Z3 [11]) is used. The logical consistency is evaluated as:

$$\text{SMT}(\Phi) = \begin{cases} 1, & \text{if } \Phi \text{ is logically consistent,} \\ 0, & \text{otherwise.} \end{cases}$$

If $\text{SMT}(\Phi) = 0$, feedback is provided to the LLM to refine the hypothesis h or the logical set Φ. This process is repeated iteratively:

$$\Phi^{(\ell+1)} = \Gamma\Big(M_\theta\big(\text{RefinePrompt}(p, \Phi^{(\ell)})\big)\Big),$$

until the solver returns SAT (i.e., $\text{SMT}(\Phi) = 1$) or a maximum number of iterations m is reached. If no consistent set is found after m iterations, conflicting statements are discarded.

Final Hypothesis Generation. Once a consistent logical set Φ is established, the final diagnostic hypothesis is produced by integrating these validated FOL premises into a coherent natural language summary. An example output might be: "*Based on the dominant blue color patterns in TDI and the discontinuous layers observed in B-mode imaging, the burn is indicative of a full-thickness injury, suggesting that surgical intervention may be required.*"

2.2 Downstream Tasks

In the downstream tasks, we combine the diagnostic hypothesis with ultrasound image classification. Each ultrasound sample x_i is composed of a tuple:

$$x_i = \left(x_i^{\mathrm{TDI}}, x_i^{\mathrm{B}}\right),$$

where x_i^{TDI} and x_i^{B} represent the raw TDI and B-mode images, respectively. These images are first converted to RGB format and then concatenated horizontally, placing the B-mode image on the left and the TDI image on the right. This results in a composite image defined as:

$$z_i \in \mathbb{R}^{H \times W \times 3}.$$

The composite RGB image z_i is then used as input to the vision-language model (VLM) classifier.

For the **binary classification** task, which distinguishes between surgery and non-surgery cases (with labels $y_i \in \{0, 1\}$), we define a classifier function:

$$g : \mathbb{R}^{H \times W \times 3} \to [0, 1],$$

such that the probability of a positive label is given by:

$$P(y_i = 1 \mid z_i) = g(z_i).$$

To incorporate the hypothesis h, we use a logical support function $\mathcal{S}(h, \Phi, y)$, which measures how well the hypothesis supports a given decision y. The final prediction is computed as:

$$\hat{y}_i = \arg \max_{y \in \{0,1\}} \Big\{ P(y \mid z_i) + \alpha \, \mathcal{S}(h, \Phi, y) \Big\},$$

where α is a hyperparameter that balances the influence of the support function.

In the case of the **fine-grained burn depth classification** task, where classes $c \in \{1, 2, \ldots, N\}$ correspond to different burn depths (with $N = 3$ in our setting), we denote the class probabilities from a multi-class classifier as $P(c \mid z_i)$. The final prediction is then given by:

$$\hat{c}_i = \arg \max_{c \in \{1,2,\ldots,N\}} \Big\{ P(c \mid z_i) + \beta \, \mathcal{S}(h, \Phi, c) \Big\},$$

where β is the associated balancing hyperparameter.

In our implementation of Logical Support Function, the VLM classifier is prompted with a query that integrates h along with the candidate class. For instance, to compute $\mathcal{S}(h, \Phi, y)$ for a candidate class y, the system may issue a prompt such as: "*Given the diagnostic hypothesis: [h] and the logical premises: [Φ], to what extent does this information support the diagnosis of [y]?*" The VLM's textual response is then mapped to a numerical score.

2.3 Classification with Hypothesis (Proposed Method)

In the proposed classification framework, the final diagnostic hypothesis h (and its corresponding reasoning, when applicable) is incorporated as a system prompt to guide the Vision-Language Model (VLM) classifier. We explore three different variants of this method:

1. Hypothesis+VLM In this variant, the VLM receives a prompt such as "What is the degree of burn depth?" or "Is this a surgery case or not?" The classification function is defined as:

$$f_{\text{VLM}} : \mathbb{R}^{H\times W\times 3} \times \mathcal{T} \to \mathcal{Y},$$

where $\mathcal{Y}$ is the label space, which could be $\{0, 1\}$ for binary classification or $\{1, 2, \ldots, N\}$ for multi-class classification. The final prediction is given by:

$$\hat{y}_i = f_{\text{VLM}}(z_i, h) = \arg\max_{y\in\mathcal{Y}} P\big(y \mid z_i, h\big).$$

2. Hypothesis+VLM with Chain-of-Thought (CoT) This variant enhances the model's reasoning capabilities by introducing a chain-of-thought (CoT) mechanism. A recursive reasoning process is used, defined as:

$$r^{(t)} = M_\theta\Big(z_i, h, r^{(1)}, r^{(2)}, \ldots, r^{(t-1)}\Big), \quad \text{for } t = 1, \ldots, T,$$

with $r^{(0)}$ initialized as an empty context. The complete chain-of-thought is represented as:

$$r = \{r^{(1)}, r^{(2)}, \ldots, r^{(T)}\}.$$

Each output $r^{(t)}$ is recursively included in the system prompt for subsequent iterations, progressively refining the prediction. The CoT-enhanced classification function is expressed as:

$$f^{\text{CoT}}_{\text{VLM}} : \mathbb{R}^{H\times W\times 3} \times \mathcal{T} \times \mathcal{R} \to \mathcal{Y},$$

where $\mathcal{R}$ is the space of chain-of-thought outputs. The final prediction is computed as:

$$\hat{y}_i = f^{\text{CoT}}_{\text{VLM}}(z_i, h, r) = \arg\max_{y\in\mathcal{Y}} P\big(y \mid z_i, h, r\big).$$

3. Hypothesis+VLM with Chain-of-Thought and Self-Consistency In the third variant, multiple candidate outputs are generated by varying sampling parameters such as temperature τ and top-p within the CoT framework. Let:

$$\{\hat{y}_i^{(1)}, \ldots, \hat{y}_i^{(K)}\}$$

denote the set of candidate predictions. The final output is obtained by aggregating these predictions, for example, using majority voting or weighted averaging:

$$\hat{y}_i = \text{Aggregate}\Big(\{\hat{y}_i^{(k)}\}_{k=1}^{K}\Big).$$

Across all these variants, the hypothesis h, and when applicable the chain-of-thought r, are integrated into the system prompt. This guides the VLM classifier toward producing more informed and interpretable predictions, further supported by the logical scoring function $\mathcal{S}(h, \Phi, \cdot)$.

3 Experiments

3.1 Dataset and Experiments Settings

Our study is based on a retrospective ultrasound dataset collected over one year at a U.S. burn center. The dataset consists of B-mode and Tissue Doppler Imaging (TDI) ultrasound scans [6] from 29 human subjects, each presenting with varying burn depths, including superficial, superficial partial-thickness (second-degree), deep partial-thickness (second-degree), and full-thickness (third-degree) burns. Ground truth labels for burn severity were obtained either through histological biopsy or, when unavailable, through expert clinical assessment and consensus from burn specialists. B-mode ultrasound was used to capture and quantify structural tissue features, while TDI provided dynamic assessments of tissue integrity. A visual example of the dataset is shown in Figure 2. To ensure high imaging quality, we selected frames marked with green-labeled TDI quality indicators, which signal proper probe pressure. This initial filtering yielded 950 ultrasound frames across all patients.

To reduce redundancy from temporally adjacent frames, we applied uniform interval sampling within each video clip, minimizing the inclusion of visually repetitive frames. After this filtering, we curated a final dataset of 324 distinct ultrasound frames. Of these, 130 frames from 15 patients were reserved for evaluation, while the remaining data were used for training purposes, including chain-of-thought and n-shot prompting strategies.

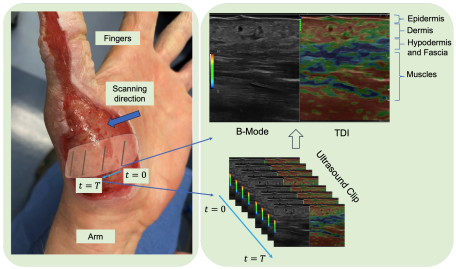

Fig. 2. An example from the burn dataset, showing B-mode and TDI ultrasound images captured from the wound site.

Implementation Details To generate diagnostic hypotheses within the KODER framework, we utilized OpenAI's `o3-mini-high` model, which is optimized for reasoning tasks. For the vision-language modeling component, we evaluated several state-of-the-art models, including OpenAI's `gpt-4o-2024-11-20`, `gpt-4o-mini-2024-07-18`, and `gpt-4-turbo-2024-04-09` [1], as well as Google's `gemini-2.0-flash` and `gemini-1.5-flash` [15].

To validate the first-order logic (FOL) premises, we employed the Z3 SMT solver [11] to ensure logical consistency. For Chain-of-Thought (CoT) reasoning [20], n-shot prompting was used, where each prompt included an ultrasound image, expert-provided explanation, and corresponding label to guide step-by-step reasoning. In the self-consistency experiments [19], we varied the VLM's temperature, top-p sampling values, and the order and number of CoT examples to analyze their impact on prediction stability. Final decisions were obtained using majority voting across outputs from multiple reasoning paths. For fine-grained burn depth classification, we implemented a two-step VLM querying process. In the first step, the model predicted whether a case was a third-degree

burn. In the second step, remaining cases were classified as either second-degree deep or second-degree superficial. This hierarchical approach yielded better performance than a single-step three-way classification, as it reduced cognitive load by narrowing the decision space at each stage.

3.2 Results

Table 1 presents a comparative analysis of various VLMs with and without the KODER framework across two tasks: Surgical Decision-Making and Fine-Grained Burn Depth classification. Overall, KODER—particularly when combined with self-consistency—significantly enhances the diagnostic performance of all evaluated models.

The best results are achieved by **GPT-4o + KODER**, which obtains an accuracy of **93%** on Surgical Decision-Making and **87%** on Burn Depth classification. In contrast, the baseline GPT-4o model without KODER shows poor performance, with only 33% and 27% accuracy, respectively, highlighting the critical role of reasoning and logical refinement in achieving high diagnostic reliability.

Other models also benefit notably from the integration of KODER. For example, **Gemini 2.0 + KODER** reaches 87% accuracy for Surgical Decision-Making, though its performance on Burn Depth classification remains moderate at 60%. Similarly, **Gemini 1.5 + KODER** demonstrates meaningful gains, achieving 80% accuracy on Surgical Decision-Making and 67% on Burn Depth, significantly outperforming its base model which only scored 60% and 47% on the two tasks, respectively.

Table 1. Performance Comparison of KODER (with self-consistency) on Surgical Decision and Fine-Grained Burn Depth.

VLM	**Surgical Decision-Making**				**Fine-Grained Burn Depth**			
	Accuracy	**F-1**	**Prec**	**Recall**	**Accuracy**	**F-1**	**Prec**	**Recall**
GPT4o+KODER	**93%**	**0.93**	**0.94**	**0.93**	**87%**	**0.87**	**0.87**	**0.87**
GPT4o	33%	0.17	0.11	0.33	27%	0.27	0.34	0.27
GPT4o-mini+KODER	80%	0.77	0.85	0.80	53%	0.42	0.35	0.53
GPT4o-mini	67%	0.67	0.69	0.67	73%	0.71	0.73	0.73
GPT4-Turbo+KODER	93%	0.93	0.94	0.93	53%	0.52	0.56	0.53
GPT4-Turbo	87%	0.87	0.87	0.87	60%	0.59	0.62	0.6
Gemini2.0+KODER	87%	0.86	0.89	0.86	60%	0.5	0.64	0.6
Gemini2.0	47%	0.41	0.79	0.47	47%	0.40	0.66	0.47
Gemini1.5+KODER	80%	0.79	0.85	0.8	67%	0.62	0.79	0.67
Gemini1.5	60%	0.5	0.42	0.6	47%	0.43	0.46	0.47

Across all configurations, the addition of KODER leads to consistent improvements in both precision and recall, indicating more reliable and balanced classification. Notably, even smaller or earlier-generation models such as `gpt-4o-mini` and `gemini-1.5` show substantial boosts in performance when paired with the KODER reasoning framework.

These results confirm that the integration of structured hypothesis generation, logical consistency checking, and chain-of-thought prompting within KODER not only enhances interpretability but also improves clinical decision-making accuracy across VLM architectures.

Overall, these results highlight the effectiveness of KODER's domain-aware representations and its ability to generate knowledge-driven hypotheses that capture the subtle nuances present in clinical ultrasound data. This leads to notable improvements in predictive performance across all evaluated models. Among them, **GPT-4o with KODER** consistently outperforms other configurations, including Gemini 2.0 and smaller-scale models. In contrast, models such as `GPT-4o Mini` and `GPT-4 Turbo` demonstrate limited capability, particularly in the fine-grained burn classification task, where distinguishing subtle variations in burn severity is essential.

Even in a zero-shot setting, the KODER framework improves overall classification accuracy to 80%, demonstrating its robustness without requiring extensive in-context examples. When combined with Chain-of-Thought (CoT) prompting, **KODER+CoT** further increases accuracy to 87%, emphasizing the value of structured intermediate reasoning in clinical decision-making. Incorporating self-consistency into this pipeline boosts performance to 93%, illustrating that multiple reasoning passes can refine predictions and yield more dependable outcomes.

A breakdown of per-class accuracy reveals that the model achieves perfect classification for second-degree superficial burns. For second-degree deep burns, the model attains 88% accuracy, with 12% misclassified as third-degree. Third-degree burns are classified with 80% accuracy, with 20% misclassified as second-degree deep. These results indicate that the model excels at detecting superficial burns but encounters more difficulty distinguishing deeper injuries, likely due to overlapping imaging characteristics among these clinically similar categories.

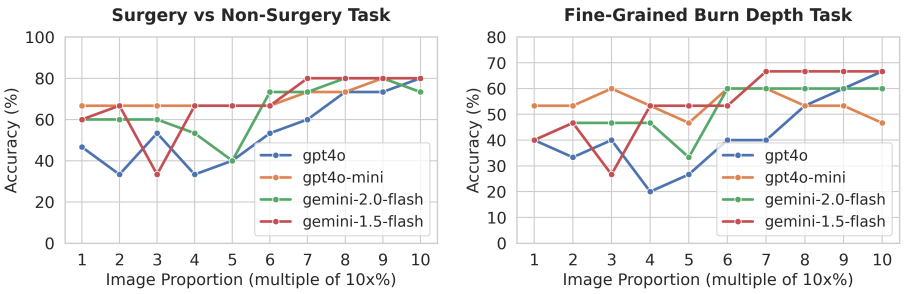

Fig. 3. Evaluation of KODER (4-shots CoT) with various image proportion.

Effect of Image Proportion on Performance To assess the robustness of KODER under varying data availability, we conducted experiments using increasing proportions of the image dataset (from right to left), ranging from 10% to 100% in 10% increments. Thus, for image proportions less than or equal to

50%, the input did not include B-mode information. The goal of this experiment was to evaluate how the amount of image input influences the performance of different VLMs when integrated with KODER.

Figure 3 shows the accuracy trends for the Surgery vs. Non-Surgery task (left) and Fine-Grained Burn Depth classification (right). Across both tasks, we observe that model (4-shots CoT) performance generally improves as more image data is included, although the rate and consistency of improvement vary across models.

In the **Surgery vs. Non-Surgery task**, `gemini-1.5-flash` demonstrates strong and stable performance, reaching its peak (approximately 80%) by the 6× image proportion and maintaining it thereafter. `gpt4o-mini` also performs consistently well across all data sizes. In contrast, `gpt4o` shows significant performance variability at lower image proportions but steadily improves with more data, eventually matching the top-performing models at the full dataset size. `gemini-2.0-flash` shows moderate improvements but fluctuates at lower proportions.

For the **Fine-Grained Burn Depth task**, the trends are more varied. `gemini-1.5-flash` again leads, achieving around 67% accuracy at the highest image proportion. `gpt4o` shows notable improvements as more images are used, especially after the 6× mark, suggesting that this model benefits from larger image contexts. `gpt4o-mini` and `gemini-2.0-flash` perform relatively well at low proportions but do not improve significantly with additional data, indicating a potential ceiling in their fine-grained classification ability under this setting.

These results suggest that while all models benefit from more image input, larger and more advanced models like `gemini-1.5-flash` and `gpt4o` scale better with increased visual context, particularly for tasks requiring fine-grained reasoning.

3.3 Ablation Study

Impact of Number CoT Shots To better understand the contribution of in-context learning, we conducted an ablation study evaluating the performance of the KODER framework across different numbers of shots (0 to 4). Table 2 reports the classification accuracy for both Surgical Decision-Making and Fine-Grained Burn Depth tasks. Overall, we observe that increasing the number of in-context examples generally improves model performance. This trend is more pronounced in the binary Surgical Decision task, where several models (e.g., GPT-4o, Gemini2.0, and GPT-4 Turbo) reach their peak accuracy (87%) by the fourth shot. GPT-4o maintains a consistently high performance across all settings, highlighting its strong reasoning capabilities even in zero-shot conditions. Notably, GPT-4o-mini also benefits from one-shot prompting, achieving an 80% accuracy that surpasses its zero-shot performance.

For the more challenging Burn Depth classification task, improvements are more gradual but still significant. GPT-4o reaches 87% accuracy at shot-4, demonstrating the advantage of progressive reasoning refinement. Gemini2.0 shows a similar pattern, peaking at 73% from a lower zero-shot baseline of 60%.

Table 2. Evaluation with number of shots. Performance (%) of KODER framework on Surgical Decision-Making and Fine-Grained Burn Depth

KODER+	Shot-0	Shot-1	Shot-2	Shot-3	Shot-4
Surgical Decision					
GPT4o	**80**	80	**80**	**80**	**87**
Gemini2.0	67	67	**87**	**80**	80
GPT4o-mini	73	**80**	73	73	73
GPT4-Turbo	73	67	67	**80**	**87**
Burn Depth					
GPT4o	**67**	67	**80**	**80**	**87**
Gemini2.0	60	**73**	73	73	73
GPT4o-mini	60	47	60	60	60
GPT4-Turbo	40	33	33	47	53

GPT-4 Turbo starts lower (40% in shot-0) but improves steadily with more examples, achieving 53% by shot-4. This suggests that while some models may struggle in zero-shot settings, they can benefit substantially from structured examples, especially in tasks requiring fine-grained differentiation.

Interestingly, performance gains plateau or slightly fluctuate for some models after two to three shots, implying diminishing returns beyond a certain point. This observation highlights the importance of balancing context complexity and example quantity in prompt design for clinical reasoning tasks.

Impact of Explanation Generation Further, we evaluate GPT-4o's performance within the KODER framework under two hypothesis generation settings. In the *Label Only* setting, the model outputs only the predicted class label. This configuration achieves an accuracy of 87%, an F1-score of 0.86, a precision of 0.89, and a recall of 0.87. In contrast, the *Explain+Label* setting requires the model to generate an explanation followed by the class label. This more structured prompting yields improved results, reaching an accuracy of 93%, an F1-score of 0.93, a precision of 0.94, and a recall of 0.93. These findings suggest that explicit explanation generation supports more accurate predictions by prompting the model to engage in deeper reasoning before outputting a label.

Qualitative Analysis Figure 4 illustrates this effect with a representative example. In the sample case, the input imaging (left) corresponds to a

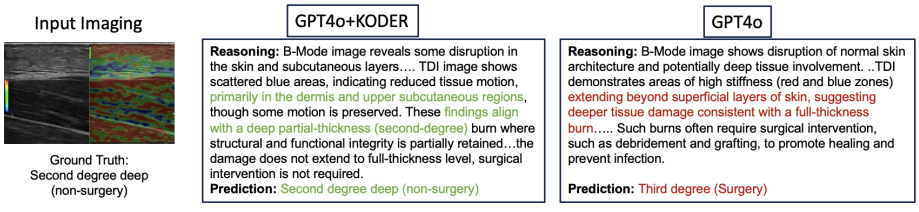

Fig. 4. In the sample results, the image input (left) reveals that GPT-4o (right) incorrectly identifies dermis layer damage as a third-degree burn rather than a second-degree deep burn. In comparison, the KODER framework (middle) demonstrates improved reasoning over the dermis layer, resulting in an accurate classification.

second-degree deep (non-surgical) burn. The **GPT-4o+KODER** model (middle) correctly classifies the burn, producing a coherent explanation that highlights dermal layer involvement and preserved structural integrity—features consistent with a deep partial-thickness injury. In contrast, the base **GPT-4o** model (right) overestimates the severity by interpreting the damage as consistent with full-thickness tissue involvement, leading to an incorrect classification of a third-degree (surgical) burn. This example reinforces the benefit of explanation-augmented reasoning in improving model reliability and interpretability in clinical tasks.

4 Conclusion

In this work, we introduced the KODER framework, a novel approach that combines domain-specific clinical knowledge with advanced reasoning techniques to improve ultrasound-based burn diagnosis. By incorporating structured reasoning through chain-of-thought prompting and enforcing consistency across multiple reasoning paths, KODER delivers substantial gains in diagnostic accuracy. For instance, in surgical decision-making, GPT-4o paired with KODER achieved up to 93% accuracy, while in fine-grained burn depth classification, it reached 87% accuracy. These results demonstrate that explicitly guiding the model to reason through its predictions leads to more accurate and interpretable outputs. KODER also showed strong performance across multiple VLMs, particularly under low-data settings, and adapted well across both binary and multi-class tasks. Overall, KODER represents a significant step toward robust, interpretable, and knowledge-guided diagnostic modeling in medical imaging applications.

Acknowledgments. This work was supported by the Office of the Assistant Secretary of Defense for Health Affairs under Award No. W81XWH-21-2-0030 and by NIH under Grant No. 5R21LM013711-02.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., OpenAI: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Bombaro, K.M., Engrav, L.H., Carrougher, G.J., Wiechman, S.A., Faucher, L., Costa, B.A., Heimbach, D.M., Rivara, F.P., Honari, S.: What is the prevalence of hypertrophic scarring following burns? Burns **29**, 299–302 (2003), publisher: Elsevier
3. Chen, X., Chi, R.A., Wang, X., Zhou, D.: Premise order matters in reasoning with large language models. In: Forty-first International Conference on Machine Learning (2024)

4. Gu, B., Desai, R.J., Lin, K.J., Yang, J.: Probabilistic medical predictions of large language models. npj Digital Medicine **7**(1), 367 (2024)
5. Guo, Y., Zeng, X., Zeng, P., Fei, Y., Wen, L., Zhou, J., Wang, Y.: Common vision-language attention for text-guided medical image segmentation of pneumonia. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 192–201. Springer (2024)
6. Ho, C.Y., Solomon, S.D.: A clinician's guide to tissue doppler imaging. Circulation **113**, e396–e398 (2006)
7. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. arXiv preprint arXiv:2306.00890 (2023)
8. Li, Q., Yan, X., Xu, J., Yuan, R., Zhang, Y., Feng, R., Shen, Q., Zhang, X., Wang, S.: Anatomical structure-guided medical vision-language pre-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 80–90. Springer (2024)
9. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26296–26306 (2024)
10. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Thirty-seventh Conference on Neural Information Processing Systems (2023)
11. de Moura, L.M., Bjørner, N.S.: Z3: An efficient smt solver. In: International Conference on Tools and Algorithms for Construction and Analysis of Systems (2008)
12. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763 (2021)
13. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI blog **1**, 9 (2019)
14. Shakeri, F., Huang, Y., Silva-Rodríguez, J., Bahig, H., Tang, A., Dolz, J., Ben Ayed, I.: Few-shot adaptation of medical vision-language models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 553–563. Springer (2024)
15. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
16. Thatcher, J.E., Squiers, J.J., Kanick, S.C., King, D.R., Lu, Y., Wang, Y., Mohan, R., Sellke, E.W., DiMaio, J.M.: Imaging techniques for clinical burn assessment with a focus on multispectral imaging. Advances in wound care **5**, 360–378 (2016), publisher: Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA
17. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
18. Tuncer, H.B., Akın, M., Çakırca, M., Erkılıç, E., Yıldız, H.F., Yastı, A.Ç.: Do pre-burn center management algorithms work? evaluation of pre-admission diagnosis and treatment adequacy of burn patients referred to a burn center. Journal of Burn Care & Research **45**(1), 180–189 (2024)
19. Wang, X., Wei, J., Schuurmans, D., Le, Q.V., Chi, E.H., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language

models. In: The Eleventh International Conference on Learning Representations (2023)
20. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
21. Zhang, J., Wang, G., Kalra, M.K., Yan, P.: Disease-informed adaptation of vision-language models. IEEE Transactions on Medical Imaging (2024)
22. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Tupini, A., Wang, Y., Mazzola, M., Shukla, S., Liden, L., Gao, J., Crabtree, A., Piening, B., Bifulco, C., Lungren, M.P., Naumann, T., Wang, S., Poon, H.: A multimodal biomedical foundation model trained from fifteen million image–text pairs. NEJM AI **2**(1) (2024). https://doi.org/10.1056/AIoa2400640