Vision-Language Reasoning for Burn Depth Assessment with Structured Diagnostic Hypotheses

Anonymous Author(s)
Affiliation
Address
email

Abstract

Ultrasound and other medical imaging data hold significant promise for burn depth assessment but remain underutilized in clinical workflows due to limited data availability, interpretive complexity, and the absence of standardized integration. Vision-language models (VLMs) have demonstrated impressive general-purpose capabilities across image and text domains, but they struggle to generalize to medical imaging modalities such as ultrasound, which are largely absent from pretraining corpora and represent a fundamentally different form of data. We present a framework for fine-grained burn depth assessment that combines digital photographs with ultrasound data, guided by structured vision-language reasoning. A central component of our method is the use of structured diagnostic hypotheses that describe clinical findings relevant to burn severity. These hypotheses can be provided by expert surgeons or automatically generated using large language models through a controlled prompting process. The reasoning process is further supported by symbolic consistency checks and chain-of-thought logic to align hypotheses with visual features, enhancing both interpretability and diagnostic performance. Our results show that the proposed method, when guided by structured reasoning, achieves higher diagnostic accuracy in burn depth assessment compared to base vision-language models without structured guidance. Importantly, the proposed system surpasses the diagnostic accuracy of expert surgeons using traditional assessment methods. This work demonstrates how multi-modal fusion and structured reasoning can enhance the explainability and accuracy of vision-language models in high-stakes medical applications.

1 Introduction

Accurate burn depth assessment is critical for determining whether a wound will heal conservatively or requires surgical intervention, such as excision and grafting. However, current clinical workflows rely heavily on subjective visual inspection and physician experience, leading to considerable inter-observer variability, delayed decision-making, and suboptimal outcomes. Even among experienced clinicians, diagnostic accuracy is estimated to range between 70% and 80%. There is a clear need for intelligent, interpretable systems that can support high-stakes clinical reasoning with greater consistency and precision.

Medical imaging data, including B-Mode ultrasound and Tissue Doppler Imaging (TDI), offer valuable physiological insights that can augment visual assessments [Ho and Solomon, 2006, Gnyawali et al., 2015, 2020]. Yet these modalities remain underutilized in burn care due to interpretive complexity, limited expert availability, and a lack of standardized integration. Annotated datasets for these modalities are scarce, and interpreting them often requires specialized knowledge that is not easily scalable. In particular, B-Mode and TDI images capture structural and perfusion-related signals that

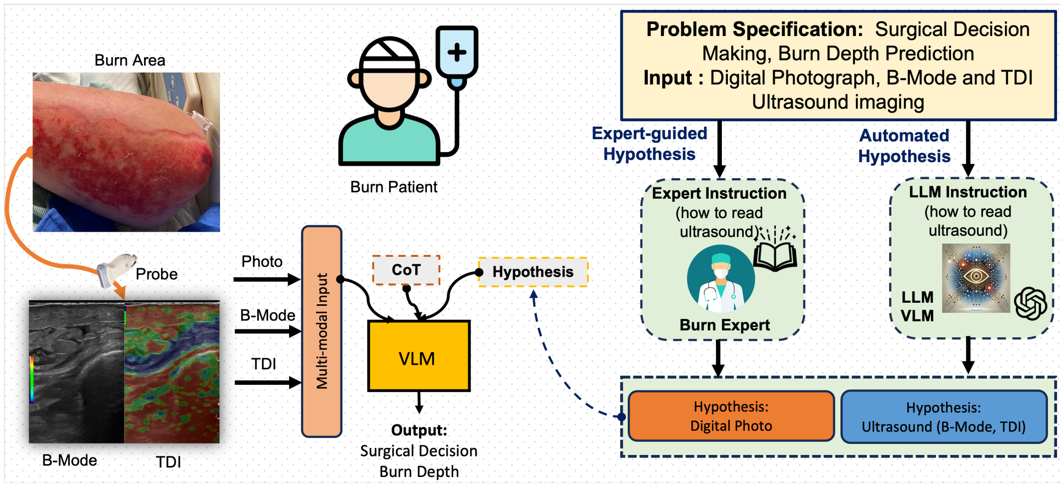

Figure 1: Overview of the proposed framework for burn depth assessment. The system takes multi-modal inputs, including digital photographs, B-mode ultrasound, and TDI ultrasound, from the burn site. Structured diagnostic hypotheses are provided either by expert surgeons or automatically generated by large-language model (LLM). These hypotheses guide the vision–language model (VLM) through chain-of-thought (CoT) reasoning to produce interpretable outputs for surgical decision making and fine-grained burn depth prediction.

are not discernible from photographs alone, making them a promising but underexploited source of diagnostic information.

At the same time, vision-language models (VLMs) have emerged as powerful tools for visual understanding and general-purpose reasoning across multimodal inputs. While these models excel in tasks such as image captioning, visual question answering, and grounded classification [Radford et al., 2019, 2021, Achiam et al., 2023, Touvron et al., 2023, Liu et al., 2023, 2024], they struggle to generalize to domain-specific data such as ultrasound, which is largely absent from their pretraining corpora and fundamentally differs from natural image distributions [Li et al., 2023, Zhang et al., 2024b, Li et al., 2024, Guo et al., 2024, Zhang et al., 2024a, Shakeri et al., 2024].

Applying VLMs to high-stakes medical imaging tasks remains challenging, particularly in scenarios where data is limited and domain expertise is essential. To address these challenges, we present a framework for fine, grained burn depth assessment that fuses digital photographs with ultrasound imaging, guided by structured vision-language reasoning. As shown in Figure 1, our method introduces structured diagnostic hypotheses, short natural language descriptions of clinically relevant features, that serve as an intermediate reasoning layer. These hypotheses can be authored by expert surgeons or automatically generated using large language models through controlled prompting strategies. We further align these hypotheses with image features using symbolic consistency checks and chain-of-thought reasoning, enabling transparent and clinically grounded predictions.

We evaluate our approach on two tasks: (i) binary classification of surgical versus non-surgical cases, and (ii) a three-way classification of burn severity: superficial partial-thickness, deep partial-thickness, and full-thickness burns. The system achieves 95% accuracy using expert-generated hypotheses and 93% with automatically generated ones—exceeding typical clinician performance and strong non-expert baselines. These improvements are consistent across multiple foundation models, including `GPT-4o`, `GPT-4 Turbo`, `Gemini 1.5`, and `Gemini 2.0`, indicating that the approach generalizes across architectures.

By integrating multi-modal fusion, structured hypothesis generation, and symbolic reasoning, this work shows how general-purpose VLMs can be adapted to support domain-specific, high-stakes diagnostic tasks. Our results demonstrate the potential for automated systems to augment clinical decision-making with expert-level accuracy and interpretability in low-data, high-risk medical environments.

2 Problem Formulation

We frame burn depth assessment as a multi-modal, hypothesis-guided visual reasoning task. Each training sample x_i is composed of three complementary imaging modalities: a digital photograph $x_i^{\mathrm{P}} \in \mathbb{R}^{H\times W\times 3}$, a B-mode ultrasound image $x_i^{\mathrm{B}} \in \mathbb{R}^{H\times W\times 3}$, and a Tissue Doppler Imaging (TDI) scan $x_i^{\mathrm{T}} \in \mathbb{R}^{H\times W\times 3}$. For each modality, we associate a modality-specific prompt c^{mod} that describes the imaging context and a structured diagnostic hypothesis h^{mod}, which may be authored by expert clinicians or generated automatically through a controlled prompting process of a large language model (LLM). The full multi-modal input is therefore:

$$\tilde{x}_i = \{(x_i^{\mathrm{P}}, c^{\mathrm{P}}, h^{\mathrm{P}}), (x_i^{\mathrm{B}}, c^{\mathrm{B}}, h^{\mathrm{B}}), (x_i^{\mathrm{T}}, c^{\mathrm{T}}, h^{\mathrm{T}})\},$$

and we define $H_i = \{h^{\mathrm{P}}, h^{\mathrm{B}}, h^{\mathrm{T}}\}$ as the set of all hypotheses tied to sample i.

Our objective is to infer clinically meaningful labels by jointly leveraging the image evidence and the structured hypotheses. For surgical decision-making, we pose a binary classification task where the ground-truth label $y_i \in \{0, 1\}$ indicates whether surgical intervention is required. The vision-language model (VLM) estimates the likelihood:

$P(y_i = 1 \mid \tilde{x}_i)$, and the final decision is computed as:

$$\hat{y}_i = \arg\max_{y\in\{0,1\}} \big[P(y \mid \tilde{x}_i) + \alpha \cdot \mathcal{S}(H_i, y)\big],$$

where $\alpha \geq 0$ weights the contribution of hypothesis alignment.

For fine-grained severity estimation, the model assigns a depth label $c_i \in \{1, \ldots, N\}$ with probabilities:

$$P(c_i = c \mid \tilde{x}_i), \quad c \in \{1, \ldots, N\},$$

and final prediction:

$$\hat{c}_i = \arg\max_{c\in\{1,\ldots,N\}} \big[P(c \mid \tilde{x}_i) + \beta \cdot \mathcal{S}(H_i, c)\big],$$

where $\beta \geq 0$ controls the influence of hypothesis-driven reasoning.

The support function $\mathcal{S}(H_i, y)$ or $\mathcal{S}(H_i, c)$ evaluates the semantic alignment between the set of diagnostic hypotheses and a candidate output, returning a scalar in $[0, 1]$. From a modeling perspective, $\mathcal{S}$ provides an auxiliary reasoning signal that can modulate predictions from the visual backbone. This is particularly valuable in medical settings where hypotheses encode domain knowledge that is otherwise difficult to capture with standard pretraining.

We operationalize $\mathcal{S}$ using chain-of-thought prompting within the VLM. For a given sample, the model is queried with a structured prompt such as: *"Given the following hypotheses from the photo, B-mode, and TDI: [. . .], how well do they support a diagnosis of full-thickness burn?"* The VLM generates a free-form reasoning trace, which is then parsed into a quantitative support score. This mapping can be implemented through a rule-based evaluation (e.g., counting agreement phrases or confidence indicators) or through a lightweight learned calibration model trained to predict alignment scores from reasoning traces.

Importantly, the support function is agnostic to the underlying VLM and is designed to be modular. It can be precomputed for a set of hypotheses or adapted online, enabling integration with a range of backbone architectures. By providing a structured, interpretable alignment signal, $\mathcal{S}$ serves as a bridge between human-readable hypotheses and the model's predictive distribution, allowing explicit incorporation of domain reasoning into both binary surgical decisions and fine-grained burn severity predictions.

3 Methodology

We propose a hypothesis-guided vision–language reasoning framework for burn depth assessment (see Figure 1). The method integrates multi-modal imaging data with structured diagnostic hypotheses, enabling a Vision–Language Model (VLM) to reason beyond raw pixel evidence and produce clinically meaningful predictions.

Input Representation and Hypothesis Construction. Each input sample x_i is composed of three complementary modalities: a digital photograph $x_i^{\mathrm{P}} \in \mathbb{R}^{H\times W\times 3}$, a B-Mode ultrasound image

$x_i^{\mathrm{B}} \in \mathbb{R}^{H \times W \times 3}$, and a Tissue Doppler Imaging (TDI) scan $x_i^{\mathrm{T}} \in \mathbb{R}^{H \times W \times 3}$. Importantly, the digital photograph and ultrasound data do not correspond frame-by-frame. Ultrasound data are acquired as video sequences, while the digital photograph is a single still image. Therefore, the photograph is first processed independently by the VLM to extract visual reasoning cues, and its resulting hypothesis and feature embedding are later combined with those derived from the ultrasound modalities for the final prediction. Practical ultrasound acquisition settings (such as frame sampling, probe parameters, and TDI configurations) are described in detail in the experimental section.

Each modality is paired with a modality-specific prompt c^{mod} that encodes acquisition context and a structured hypothesis h^{mod} that describes modality-specific diagnostic cues. Hypotheses can be sourced from two streams: (i) *expert-guided*, provided directly by experienced burn surgeons, and (ii) *automated*, generated by a large language model M_θ given modality instructions: $h^{\mathrm{mod}} = M_\theta(c^{\mathrm{mod}})$, yielding a hypothesis set $H_i = \{h^{\mathrm{P}}, h^{\mathrm{B}}, h^{\mathrm{T}}\}$. The fused input is: $\tilde{x}_i = \big\{(x_i^{\mathrm{P}}, c^{\mathrm{P}}, h^{\mathrm{P}}), (x_i^{\mathrm{B}}, c^{\mathrm{B}}, h^{\mathrm{B}}), (x_i^{\mathrm{T}}, c^{\mathrm{T}}, h^{\mathrm{T}})\big\}$.

Support Function and Reasoning. To incorporate structured reasoning, we define a support function $\mathcal{S}(H_i, y)$ that quantifies how well the hypotheses support a candidate decision y: $\mathcal{S} : H_i \times \mathcal{Y} \to [0, 1]$. This is implemented via chain-of-thought prompting of the VLM. For example, the model may be queried with: *"Given the following hypotheses from the photo, B-mode, and TDI: [. . .], how well do they support a diagnosis of full-thickness burn?"* The reasoning trace from the VLM is then parsed and converted into a scalar support score, using either rule-based evaluation (e.g., detecting agreement indicators) or a lightweight learned calibration model.

Prediction with Hypothesis-Guided Support. The VLM predicts label probabilities from the fused input: $P(y \mid \tilde{x}_i), \quad y \in \mathcal{Y}$, where $\mathcal{Y} = \{0, 1\}$ for surgical decision or $\mathcal{Y} = \{1, \ldots, N\}$ for multi-class burn severity. Final predictions integrate hypothesis support: $\hat{y}_i = \arg\max_{y \in \mathcal{Y}} \Big[P(y \mid \tilde{x}_i) + \lambda \cdot \mathcal{S}(H_i, y)\Big]$, with $\lambda \geq 0$ controlling the influence of hypothesis-guided reasoning.

Algorithm. The overall procedure is summarized in Algorithm 1. This algorithm shows how multi-modal inputs and expert- or LLM-generated hypotheses are combined within the VLM and refined through the support function to yield the final decision. By explicitly modeling reasoning through diagnostic hypotheses, the framework enables the VLM to combine information from digital photographs and ultrasound imaging, even though these inputs are not temporally aligned frame-by-frame. This design allows the system to incorporate expert knowledge or automatically generated insights, leading to interpretable and accurate predictions for both surgical decision making and fine-grained burn depth classification.

Algorithm 1: Hypothesis-Guided Burn Depth Assessment

Input: Multi-modal inputs for i: $x_i^{\mathrm{P}}, x_i^{\mathrm{B}}, x_i^{\mathrm{T}}$;
Prompts $c^{\mathrm{P}}, c^{\mathrm{B}}, c^{\mathrm{T}}$;
Hypothesis source (expert or LLM M_θ);
VLM; parameter λ.
Output: $\hat{y}_i \in \mathcal{Y}$.
1. Hypothesis Generation:
foreach $\mathrm{mod} \in \{\mathrm{P}, \mathrm{B}, \mathrm{T}\}$ **do**
    if *expert* **then**
        $h^{\mathrm{mod}} \leftarrow$ expert hypothesis;
    else
        $h^{\mathrm{mod}} \leftarrow M_\theta(c^{\mathrm{mod}})$;
$H_i \leftarrow \{h^{\mathrm{P}}, h^{\mathrm{B}}, h^{\mathrm{T}}\}$;
2. Fuse Inputs:
$\tilde{x}_i \leftarrow \{(x_i^{\mathrm{mod}}, c^{\mathrm{mod}}, h^{\mathrm{mod}})\}_{\mathrm{mod}}$;
3. VLM Prediction:
Compute $P(y \mid \tilde{x}_i)$;
4. Support Scoring:
foreach $y \in \mathcal{Y}$ **do**
    Query VLM with H_i, y;
    Score $s_y \leftarrow \mathcal{S}(H_i, y)$;
5. Final Decision:
$\hat{y}_i \leftarrow \arg\max_y \big[P(y \mid \tilde{x}_i) + \lambda \cdot s_y\big]$;
return $\hat{y}_i$;

Automated Hypothesis Generation. To enable structured reasoning without relying exclusively on human expertise, we introduce an automated hypothesis generation module (see Figure 2). This component is designed to transform clinical knowledge and experimental details into machine-readable hypotheses that guide the VLM in interpreting ultrasound data for burn depth prediction.

The process begins by constructing a prompt that fuses two forms of textual context: the experimental setup and the clinical interpretation of imaging cues. Let $\mathcal{D}_{\mathrm{exp}}$ represent descriptive details of the imaging modalities (e.g., "TDI provides color-coded velocity maps; B-mode offers structural tissue layers"), and $\mathcal{D}_{\mathrm{clin}}$ capture clinical heuristics (e.g., "dominant blue regions in TDI and disrupted layers in B-mode correlate with full-thickness burns"). These are concatenated using a PromptBuilder

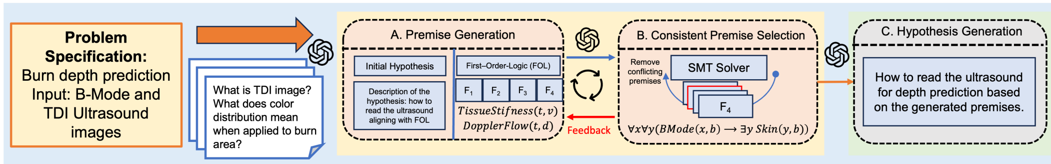

Figure 2: **Automated hypothesis generation pipeline.** Given the problem specification (burn depth prediction using B-Mode and TDI ultrasound), the system first constructs an input prompt by combining experimental descriptions with clinical context. (A) The language model generates initial hypotheses and corresponding first-order logic (FOL) premises describing how to interpret ultrasound patterns. (B) An SMT solver iteratively filters out conflicting premises and enforces logical consistency, providing feedback to refine the generated rules. (C) The validated premises are summarized into a final natural-language hypothesis that guides the vision-language model in subsequent burn depth assessment.

function: $p = \text{PromptBuilder}(\mathcal{D}_{\text{exp}}, \mathcal{D}_{\text{clin}})$, which yields a structured query that supplies the language model with sufficient background knowledge.

Given this prompt p, a large language model M_θ generates both an initial natural-language hypothesis h and a set of first-order logic (FOL) premises $\Phi = \{\phi_1, \phi_2, \ldots, \phi_K\}$ that encode specific diagnostic rules. Sampling parameters such as temperature and top-p nucleus sampling are varied to encourage diverse candidate rules.

To validate the first-order logic (FOL) premises, we employed the Z3 SMT solver [de Moura and Bjørner, 2008] to ensure logical consistency. The solver iteratively removes contradictions and provides feedback to refine the generated premises. This cycle continues until a logically consistent set is obtained or a maximum iteration threshold is reached. Conflicting statements are pruned, and the remaining validated premises are then summarized into a final natural-language hypothesis. A typical output might be: *"Based on the presence of dominant blue regions in TDI and discontinuous layers in B-mode, the burn is indicative of full-thickness injury and may require surgical intervention."*

By integrating this automated pipeline into the multi-modal reasoning framework, the system can dynamically generate domain-specific guidance without requiring manual annotations or handcrafted rules, significantly improving interpretability and diagnostic accuracy.

4 Experiments

4.1 Dataset and Experimental Setup

We evaluate our approach on a retrospective dataset collected over a one-year period at a major U.S. burn treatment center. To our knowledge, this is the **first dataset** to pair **Tissue Doppler Imaging (TDI)** and **B-Mode ultrasound** for **burn depth assessment**, enabling multi-modal reasoning beyond traditional RGB imagery. The dataset includes ultrasound recordings from 29 patients with clinically verified burn injuries spanning superficial, superficial partial-thickness, deep partial-thickness, and full-thickness (third-degree) burns. Ground-truth depth labels were determined via histological biopsy when available (5 cases) or established through consensus among board-certified burn surgeons.

Each ultrasound sample contains both B-Mode frames, capturing structural echogenicity, and TDI frames, encoding perfusion-sensitive velocity information via pseudo-color. To ensure data reliability, we retained only TDI frames flagged as high-quality by the acquisition system (indicated by green diagnostic markers ensuring optimal probe placement and coupling). From the raw sequences, 950 high-quality frames were extracted and then uniformly sampled at fixed intervals to reduce redundancy and maximize scene diversity, yielding 324 unique frames for downstream analysis. We hold out 130 frames from 15 subjects for evaluation, while the remainder are used to construct few-shot prompts, chain-of-thought demonstrations, and calibration examples.

It is important to note that digital photographs of the burn sites, captured at bedside, do not align frame-by-frame with ultrasound sequences. Photographs are single still images processed independently by the VLM, and their reasoning outputs are later fused with ultrasound-derived cues.

Ultrasound acquisition parameters (e.g., probe frequency, TDI velocity ranges) follow clinical best practices and are detailed in the experimental section of the supplementary material. Representative samples are shown in Figure 3, highlighting the complementary information captured by photographs and ultrasound modalities.

Hypothesis Generation and Vision–Language Models. For automated diagnostic hypotheses, we use OpenAI's `o3-mini-high`, a compact LLM optimized for symbolic reasoning and logical chaining. These hypotheses complement expert-provided ones within our framework. Vision–language reasoning tasks are carried out on multiple foundation models, including `gpt-4o`, `gpt-4o-mini`, `gpt-4-turbo`, `gemini-2.0-flash`, and `gemini-1.5-flash`. These models were selected for their demonstrated multi-modal reasoning capabilities, low latency, and compatibility with structured prompts that integrate visual evidence and text.

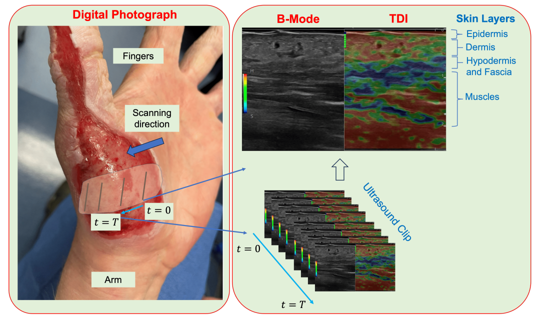

Figure 3: Representative samples from the dataset. Left: digital photographs of burn wounds. Right: paired ultrasound images, with B-Mode showing tissue structure and TDI visualizing perfusion.

All experiments are conducted in zero-shot or few-shot configurations unless otherwise stated. We benchmark our method by comparing predictions guided by expert-crafted hypotheses against those guided by automatically generated hypotheses, under identical input conditions and prompt scaffolds, to isolate the impact of structured reasoning on burn depth assessment.

4.2 Experimental Setup

We evaluate our framework across three complementary experimental settings. Each setting is designed to isolate specific factors in diagnostic performance by varying the input modalities, the classification granularity, and the source of diagnostic hypotheses. This controlled design allows us to analyze the contribution of each component in a systematic manner.

1. Binary Surgical Decision with Ultrasound (Expert vs. Automated Hypotheses). The first setting targets a binary decision task: determining whether surgical intervention is required. Inputs are limited to ultrasound data, comprising both B-Mode and Tissue Doppler Imaging (TDI). In the expert-guided variant, board-certified burn surgeons manually reviewed the ultrasound frames and produced structured diagnostic hypotheses informed by procedural knowledge and anatomical cues. TDI frames were spatially cropped using landmarks visible in the corresponding B-Mode scans, ensuring that hypotheses addressed clinically meaningful regions. To evaluate automated reasoning, the same ultrasound inputs were provided to a language model with modality-specific prompts. In this configuration, the model generated diagnostic hypotheses without any expert annotations or region-of-interest cues. This setting provides a direct comparison between expert-derived and LLM-generated reasoning when using ultrasound alone.

2. Fine-Grained Burn Classification with Ultrasound (Automated Hypotheses). The second setting assesses fine-grained classification performance using ultrasound data only, guided exclusively by automatically generated hypotheses. The task is defined over three clinically significant categories: second-degree superficial, second-degree deep, and third-degree burns. First-degree burns are excluded because they are seldom represented in hospital ultrasound workflows. This experiment probes the system's ability to differentiate subtle structural and perfusion patterns that separate intermediate burn types—cases that are difficult to resolve using image evidence alone. The use of LLM-generated hypotheses allows us to measure how structured reasoning impacts fine-grained predictions.

3. Fine-Grained Classification with Photographs and Ultrasound (Expert Hypotheses). The third setting extends the analysis to a multi-modal scenario, combining bedside digital photographs with ultrasound inputs. In this setting, hypotheses are generated exclusively by experienced burn surgeons. The classification space includes first-, second-, and third-degree burns, with superficial and deep second-degree categories merged into a single class. This merging reflects common practice in triage and telemedicine contexts, where photographs may be taken in non-clinical environments and used for remote assessments. This experiment highlights the value of photographs as a complementary modality and demonstrates the framework's ability to integrate heterogeneous inputs with expert-guided reasoning, yielding interpretable predictions across the full spectrum of burn severity.

Implementation Details: The final classification results of our method, when guided by structured diagnostic hypotheses, are obtained through a self-consistency [Wang et al., 2023] strategy combined with CoT reasoning [Wei et al., 2022]. For each input, the VLM is queried multiple times with different sampling parameters. The temperature values are sampled within the range of 0.5 to 1 and the top-p values are sampled within the range of 0.5 to 1.0. In addition, the order and the subset of CoT exemplars in the prompt are permuted to encourage diverse reasoning paths. These multiple outputs are aggregated, and the final prediction is determined by majority voting across all generated responses.

4.3 Results

We report results across binary surgical decision tasks, fine-grained burn depth classification, and multi-modal fusion analyses. All experiments compare baseline VLMs to their hypothesis-guided counterparts to quantify the impact of structured reasoning.

Table 1: Surgical decision performance using expert-guided and automated hypotheses across VLMs.

Method	Accuracy	F1	Prec	Recall
Expert Hypothesis	**95%**	**0.95**	**0.94**	**1.00**
GPT-4o + Auto Hypothesis	**93%**	**0.93**	**0.94**	**0.93**
GPT-4o (Base)	33%	0.17	0.11	0.33
GPT-4o-mini + Auto Hypothesis	80%	0.77	0.85	0.80
GPT-4o-mini (Base)	67%	0.67	0.69	0.67
GPT-4 Turbo + Auto Hypothesis	**93%**	**0.93**	**0.94**	**0.93**
GPT-4 Turbo (Base)	87%	0.87	0.87	0.87
Gemini 2.0 + Auto Hypothesis	87%	0.86	0.89	0.83
Gemini 2.0 (Base)	47%	0.41	0.79	0.47
Gemini 1.5 + Auto Hypothesis	80%	0.79	0.85	0.80
Gemini 1.5 (Base)	60%	0.50	0.42	0.60

Surgical vs. Non-Surgical Classification. Table 1 summarizes performance on the binary task of determining surgical necessity using ultrasound inputs. Expert-written hypotheses provide the strongest guidance, yielding **95% accuracy**, an F1-score of **0.95**, precision of **0.94**, and perfect recall. These results reflect close clinical alignment and serve as a reference upper bound.

Automated hypothesis generation substantially improves outcomes for all VLMs. For example, `GPT-4o` and `GPT-4 Turbo` paired with automatically generated hypotheses both achieve **93% accuracy** and an F1-score of **0.93**, approaching expert performance. In contrast, the base `GPT-4o` without hypothesis guidance reaches only **33% accuracy** and an F1-score of **0.17**, highlighting the limitations of direct image-to-text reasoning in high-stakes settings.

Smaller models benefit as well: `GPT-4o-mini` improves from 67% to 80% accuracy, `Gemini 1.5` from 60% to 80%, and `Gemini 2.0` from 47% to 87%. These gains demonstrate that structured reasoning provides a consistent boost in diagnostic alignment across diverse architectures.

Fine-Grained Burn Depth Classification with Automated Hypotheses.

Table 2: Fine-grained burn depth classification with automated hypotheses across VLMs.

Method	Accuracy	F1	Prec	Recall
GPT-4o + Auto Hypothesis	**87%**	**0.87**	**0.87**	**0.87**
GPT-4o (Base)	27%	0.27	0.34	0.27
GPT-4o-mini + Auto Hypothesis	53%	0.42	0.53	0.53
GPT-4o-mini (Base)	73%	0.71	0.73	0.73
GPT-4 Turbo + Auto Hypothesis	53%	0.52	0.56	0.53
GPT-4 Turbo (Base)	60%	0.59	0.62	0.60
Gemini 2.0 + Auto Hypothesis	60%	0.50	0.64	0.60
Gemini 2.0 (Base)	47%	0.46	0.60	0.47
Gemini 1.5 + Auto Hypothesis	67%	0.62	0.79	0.67
Gemini 1.5 (Base)	47%	0.43	0.46	0.47

Table 2 reports performance for three-class burn depth classification (first-, second-, and third-degree) using only ultrasound inputs with automated hypotheses. `GPT-4o` achieves the best results, with **87% accuracy** and balanced precision, recall, and F1-score, all at **0.87**. This is a substantial improvement over the base model's 27% accuracy, underscoring the value of explicit reasoning for fine-grained tasks.

Other VLMs also benefit from hypothesis guidance. `Gemini 1.5` improves from 47% to 67% accuracy, and `Gemini 2.0` improves from 47% to 60%. Interestingly, the base versions of `GPT-4 Turbo` and `GPT-4o-mini` perform competitively or slightly better than their hypothesis-augmented counterparts on this task, suggesting that certain architectures may already encode sufficient priors for moderate-granularity distinctions. Nonetheless, the overall trend shows that hypothesis-driven reasoning improves performance in challenging, domain-specific classification.

Effect of Multi-Modal Fusion. When using only digital photographs, the model performs well for superficial injuries, correctly identifying 83.3% of first-degree burns and 76.9% of second-degree burns. However, it struggles significantly with third-degree burns, correctly identifying only 14.3% of those cases. Incorporating TDI ultrasound features dramatically improves deep burn recognition, achieving 100% correct identification for third-degree burns while maintaining stable performance for first- and second-degree categories.

Table 3: Per-class performance comparison: digital photographs only vs. multi-modal input (photographs + TDI ultrasound). AUROC values include 95% confidence intervals, and correct classification rates are reported as percentages.

Burn Class	Setting	AUROC	95% CI	Correct (%)
1st-degree	Photo only	0.91	0.80–0.98	83.3%
	Multi-modal	0.97	0.91–1.00	83.3%
2nd-degree	Photo only	0.88	0.75–0.95	76.9%
	Multi-modal	0.96	0.90–1.00	76.9%
3rd-degree	Photo only	0.62	0.40–0.80	14.3%
	Multi-modal	1.00	1.00–1.00	100.0%

A detailed comparison of these results is provided in Table 3. For first-degree burns, the AUROC improves from 0.91 (95% CI: 0.80–0.98) with photographs alone to 0.97 (95% CI: 0.91–1.00) with multi-modal input, while maintaining the same correct classification rate of 83.3%. For second-degree burns, the AUROC increases from 0.88 (95% CI: 0.75–0.95) to 0.96 (95% CI: 0.90–1.00), again with a stable correct classification rate of 76.9%. The most striking improvement is observed for third-degree burns, where AUROC jumps from 0.62 (95% CI: 0.40–0.80) to 1.00 (95% CI: 1.00–1.00), with the correct classification rate rising from 14.3% to 100.0%.

These results demonstrate that adding ultrasound data yields measurable gains in discrimination ability across all classes, particularly for third-degree burns where structural and perfusion information is essential for reliable identification. The stable performance on less severe classes shows that integrating additional modalities does not degrade recognition for easier cases, while dramatically improving outcomes for clinically critical deep burns.

Qualitative Impact of Chain-of-Thought Reasoning. CoT reasoning plays a pivotal role in bridging raw visual evidence and clinically meaningful interpretation. To illustrate this, we analyzed representative cases processed by GPT-4o under our proposed framework (see Figure 4). The qualitative behavior reveals how step-by-step reasoning enhances both interpretability and predictive reliability.

In one challenging case, the model incorrectly predicts a third-degree burn with high confidence. Its internal reasoning shows that it detected a dominant blue region in the TDI input and mapped this pattern directly to hypodermal involvement. While blue dominance often signals tissue stiffness, the spatial distribution in this instance was confined to superficial layers and should not have triggered a full-thickness classification. The error highlights that even with structured reasoning, models may overgeneralize cues without nuanced spatial understanding. Importantly, because the CoT output explicitly described this reasoning, the source of error is transparent, offering actionable insight for refinement.

In contrast, another case demonstrates the intended use of CoT reasoning. Here, the model accurately classifies a non-third-degree burn and articulates a reasoning chain that aligns with clinical expectations. It systematically identifies relevant tissue layers, examines color gradients in the TDI scan, and concludes that no dominant blue signal extends beyond the dermis. This structured narrative not only supports the correctness of the prediction but also exposes the underlying rationale in terms that are interpretable by clinicians. These examples underscore the value of incorporating chain-of-thought reasoning in multimodal diagnostic pipelines. Rather than producing opaque predictions, the model outputs a reasoning trace that contextualizes its decision process, enabling experts to evaluate, trust, and, when necessary, challenge the system's outputs. This level of interpretability is particularly critical for deployment in high-stakes medical settings, where explainable errors and traceable successes both contribute to system validation and continuous improvement.

4.4 Discussion

A central finding of our experiments is the significant performance gap between base VLMs and their hypothesis-guided counterparts. In zero-context conditions, base models like GPT-4o often misinterpret critical TDI patterns, leading to incorrect predictions. For instance, blue dominance in TDI, which in burn imaging indicates high tissue stiffness and often correlates with deep dermal or full-thickness burns, was frequently misunderstood by the base model as benign. Without domain-specific guidance, GPT-4o sometimes associated red or green hues with stiffness and deep injury, directly contradicting the clinical interpretation of TDI color codes. These errors explain the poor baseline performance, with accuracy dropping to around 33% for surgical decision tasks when no contextual information was provided.

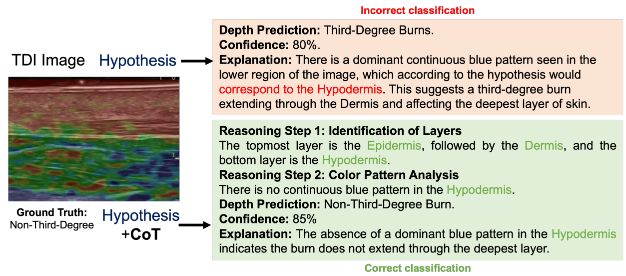

Figure 4: Qualitative examples from GPT-4o with hypothesis guidance. Top: false positive with misaligned reasoning. Bottom: correct classification with consistent reasoning.

The structured reasoning approach introduced in this work addresses these limitations by generating task-specific hypotheses that explicitly link visual patterns to clinical concepts. By providing models with contextual grounding, for example, instructing that "a dominant blue pattern in TDI suggests tissue stiffness and deeper injury", the framework enables VLMs to focus on clinically relevant features. This mechanism is particularly effective for stronger image-language models such as GPT-4o and GPT-4 Turbo, which are better able to leverage the reasoning cues. Smaller models also benefit, though to a lesser extent, due to their reduced capacity for complex multimodal reasoning.

Beyond performance improvements, the proposed framework offers practical advantages for real-world deployment. It is model-agnostic and can be integrated into any VLM workflow capable of handling multimodal inputs and textual outputs. Unlike conventional CNN- or ViT-based pipelines, which often require large-scale domain-specific pretraining, our approach relies on lightweight prompt engineering and logical hypothesis generation. This design is especially attractive for specialized domains like burn ultrasound, where large annotated datasets are scarce. Furthermore, the framework produces interpretable, text-based explanations alongside predictions, an important requirement for clinical adoption and trust. By combining minimal data requirements, improved reasoning capabilities, and interpretability, this work establishes a foundation for integrating ultrasound into broader burn assessment protocols and encourages future multi-center data collection efforts.

5 Conclusion

We introduced a vision–language framework for burn depth assessment that integrates digital photographs and ultrasound modalities with structured diagnostic reasoning. The framework incorporates both expert-authored and automatically generated hypotheses, enabling large vision–language models to interpret underrepresented imaging modalities such as B-mode and TDI ultrasound. An automated hypothesis generation module, coupled with logical consistency verification using an SMT solver, produces domain-specific reasoning instructions without requiring extensive manual annotation. Extensive experiments demonstrate that hypothesis-guided reasoning significantly improves performance compared to base VLMs. Our approach achieves up to **95%** accuracy on binary surgical decision tasks and **87%** accuracy on three-class burn depth classification, with high AUROC values across all classes. Multi-modal fusion further enhances performance, achieving higher correct identification of third-degree burns while maintaining stable accuracy on less severe cases. Qualitative analysis shows that chain-of-thought reasoning exposes the decision process, yielding interpretable predictions and revealing sources of errors. These results highlight that structured reasoning, combined with multi-modal inputs, can adapt general-purpose VLMs to high-stakes clinical tasks. The proposed framework offers both improved diagnostic performance and interpretable outputs, establishing a foundation for trustworthy deployment of vision–language systems in medical imaging workflows.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Leonardo Mendonça de Moura and Nikolaj S. Bjørner. Z3: An efficient smt solver. In *International Conference on Tools and Algorithms for Construction and Analysis of Systems*, 2008.

Surya C Gnyawali, Kasturi G Barki, Shomita S Mathew-Steiner, Sriteja Dixith, Daniel Vanzant, Jayne Kim, Jennifer L Dickerson, Soma Datta, Heather Powell, Sashwati Roy, et al. High-resolution harmonics ultrasound imaging for non-invasive characterization of wound healing in a pre-clinical swine model. *PloS one*, 10:e0122327, 2015. Publisher: Public Library of Science.

Surya C Gnyawali, Mithun Sinha, Mohamed S El Masry, Brian Wulff, Subhadip Ghatak, Fidel Soto-Gonzalez, Traci A Wilgus, Sashwati Roy, and Chandan K Sen. High resolution ultrasound imaging for repeated measure of wound tissue morphometry, biomechanics and hemodynamics under fetal, adult and diabetic conditions. *PLoS One*, 15:e0241831, 2020.

Yunpeng Guo, Xinyi Zeng, Pinxian Zeng, Yuchen Fei, Lu Wen, Jiliu Zhou, and Yan Wang. Common vision-language attention for text-guided medical image segmentation of pneumonia. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 192–201. Springer, 2024.

Carolyn Y Ho and Scott D Solomon. A clinician's guide to tissue doppler imaging. *Circulation*, 113: e396–e398, 2006.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023.

Qingqiu Li, Xiaohan Yan, Jilan Xu, Runtian Yuan, Yuejie Zhang, Rui Feng, Quanli Shen, Xiaobo Zhang, and Shujun Wang. Anatomical structure-guided medical vision-language pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 80–90. Springer, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1:9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763, 2021.

Fereshteh Shakeri, Yunshi Huang, Julio Silva-Rodríguez, Houda Bahig, An Tang, Jose Dolz, and Ismail Ben Ayed. Few-shot adaptation of medical vision-language models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 553–563. Springer, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Jiajin Zhang, Ge Wang, Mannudeep K Kalra, and Pingkun Yan. Disease-informed adaptation of vision-language models. *IEEE Transactions on Medical Imaging*, 2024a.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1), 2024b. doi: 10.1056/AIoa2400640.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims are supported by a mathematical formalization of the problem and empirical validation using a relevant dataset.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of the method are discussed in the Results section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: Not applicable.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Implementation details and pre-trained models are described in the Experiments section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The code and dataset will be made publicly available upon paper acceptance to ensure reproducibility and support future research.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Relevant details are provided in the Experiments section and the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: Explanations for the analysis are provided where necessary.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Inference API access details for the pre-trained models are provided to support reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research was conducted in accordance with all aspects of the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: No immediate or apparent societal risks arise from this research, and its implications are discussed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The dataset will be released with safeguards in place.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Appropriate citations are given.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: New dataset is discussed in details.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.