Critical Algorithms for Medical Triage: A Pragmatic Approach to Contextual Diagnostic Segmentation

James Emilian

Master of Science in Biomedical Engineering
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Advisor:
Prof. John Galeotti

. . . . . .

BME Faculty Reader:
Prof. Steven Chase

. . . . . .

Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science - Biomedical Engineering (Research)
December 2024

Contents

1 Introduction

Disasters are often unpredictable and demand immediate attention, especially in mass casualty incidents like battlefields or plane crashes, where large numbers of victims in inaccessible areas need urgent assessment. In such scenarios, first responders face significant constraints in assessing casualties, prioritizing critical cases, and providing life-saving interventions (LSIs). Some disasters also create hazardous environments that delay responder entry. A robotic system capable of entering dangerous situations could enable paramedics to extend their capabilities, acting as a force multiplier for medical triage and rescue efforts. In disaster zones where the scale and number of casualties far exceed the number of skilled medics, such a system can provide rapid initial assessments, which trained human medics can then use to quickly pinpoint those requiring LSIs. These systems are particularly valuable in disaster zones where the scale of casualties far exceeds the availability of skilled medics. In response to these needs, DARPA recently created their Triage Grand Challenge (i.e. the **DARPA Triage Challenge**, **DTC**), which inspired much of this present work. The DTC requires teams to deploy a fleet of robots that can enter a mass-casualty event, and in a non-contact manner sense multiple health status parameters accurately.

This thesis focuses on severe hemorrhage detection. We begin with (i) a review of the models that were tested/evaluated, (ii) performance evaluation of the models deployed during the DTC Year 1, and (iii) an overview of resulting research problems we are pursing to build a better system for DTC Year 2.

Though current literature is rich in specialized solutions to estimate vital signs like heart rate and respiratory rate in controlled environments, there has been no attempt prior to this challenge to deploy a diverse set of AI models at the edge, embodied in an active robotic system for the goal of sensing varied health parameters for medical triage. A key contribution of this work is an investigation into how existing models can be adapted and deployed for severe hemorrhage detection to perform (somewhat) accurate predictions even in out-of-distribution (OOD) settings. Another contribution is a clear breakdown of the observed shortcomings of current AI approaches for our domain-specific evaluation and field testing experiments.

2 Related Work

Early segmentation methods were primarily based on classical algorithms, such as thresholding, region growing, and edge detection, which laid the groundwork for more sophisticated approaches [1]. Notable early works include the use of graph-cut methods for segmenting anatomical structures, which demonstrated the potential of computational techniques in medical imaging [2]. These foundational techniques paved the way for the introduction of deep learning models, which revolutionized the field. The introduction of U-Net [3] in 2015 marked a pivotal moment in medical image segmentation, as it provided a robust architecture specifically designed for accurate biomedical image analysis with a vast reduction in the training dataset sizes required. U-Net's encoder-decoder structure, characterized by skip connections that merge high-resolution features with low-resolution features, significantly improved segmentation accuracy and efficiency. Various versions of the U-Net have been successful in a variety of tasks [4], [5] - including vertebral disc localization from MRI data [6] and liver and tumor segmentation from CT data [7]. Variants of the U-Net fused with transformers have been proven to be effective at medical image segmentation tasks as well ([8], [9]). The U-Net model's success spurred the development of various derivatives, including the Feature Pyramid Network (FPN) [10], which enhances multi-scale feature extraction and generalizes to both open-world object detection as well as segmentation tasks. Until the advent of large vision transformer based object detection models, U-Net and FPN models remained state-of-the-art. The FPN, with its reduced need for training data, may have been underutilized outside of some corners of medical imaging literature ([11], [12]), novel medical question answering setups ([13]), and even niche tasks such as coral polyp segmentation [11]. In our study, we compare the standard FPN to the standard U-Net, and show that it is objectively superior to standard U-Net in our task.

The need for generalizability ushers in the topic of foundation models - large models that are trained on vast corpora of data and hold rich representations of the world and/or the domain they were trained on. Recently, there has been a lot interesting work in biomedical image segmentation that build on foundation models. This includes a novel approach for joint segmentation and recognition across multiple modalities [14], cases of finetuning SAM for improved medical image segmentation [15],

amongst other works that point to the generalist capabilities of vision-language foundation models ([16], [17]). However, the task of hemorrhage detection is not a typical biomedical segmentation problem; given the open world nature of the expected challenge environments, a model with the ability to perform open-vocabulary detection and segmentation is essential. Given that, it is interesting to note the advent of a novel Grounded-SAM (GSAM) model [18] - a model that cascades the open-vocabulary object/instance detection abilities of Grounding-DINO [19] with the universal segmentation abilities of Segment Anything Model (SAM) [20]. It will be useful to deploy it as a benchmark model given its powerful segmentation abilities - and indeed, we have done so in this work.

Despite the plethora of such models available, this challenge forced us to tackle a rather unique problem - of having to address an in-the-wild segmentation task which, due to its inherently NSFW nature, posed two distinct problem - (i) lack of publicly available datasets, and (ii) high costs of data gathering and annotation. As far as the author is aware, there are multiple works tackling internal hemorrhage with models trained on clinical data ([21], [22], [23], [24])- but no work in current literature has attempted to detect external severe hemorrhage; and in that regard, this work is truly first of its kind.

The need for deployment on edge device also belies the need for a small, efficient model capable of running on devices like the NVIDIA Jetson Orin - requiring the right tradeoff between performance and computational efficiency. Furthermore, the model must be trainable with limited datasets, a common constraint in medical imaging due to the high costs associated with data annotation. Additionally, the ability to generalize across varying conditions while maintaining a low false positive rate is crucial for clinical applicability. Thus, in this thesis, the exploration of both small models like FPN and U-Net, alongside larger models such as GroundedSAM, leads to a more comprehensive framework than the ones inferable from existing literature, for addressing the complexities of hemorrhage detection. While existing research has laid a strong foundation in segmentation capabilities, our work takes a highly pragmatic approach of attempting to rigorously test and identify the best pipeline for the uniquely out-of-distribution (OOD) task of hemorrhage detection. By illustrating the said model comparisons, we throw light on the appropriateness of varied models across varied contexts of biomedical AI.

3 Deployment Context

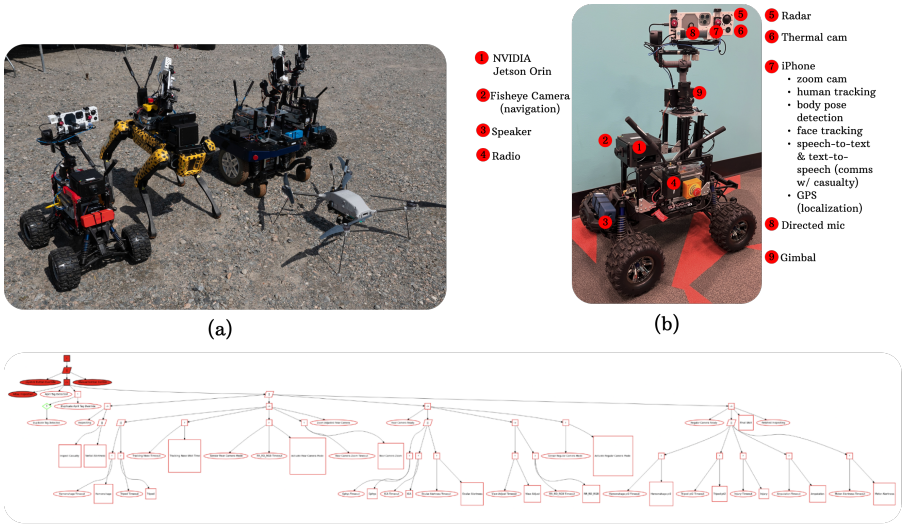

Figure 1: (a) Team Chiron robotic fleet, (b) details of payload components, and (c) the final behavior tree

Before going into the details of the models developed, it it worthwhile to look at the deployment details of the DARPA Triage Challenge. Presented with a simulated set of mass casualty incidents,

the robots are teleoperated to navigate and find casualties - upon finding a casualty, the system autonomously triggers a sequence of algorithms (see Fig. 1 (c)) to assess vitals and critical parameters of the simulated casualties. The predicted values then have to be reported to DARPA's central server, wherein the predicted values are compared to pre-defined ground truths and scored. The ground truths for vitals (heart rate and respiratory rate) are live-streamed from the casualty. See Figure 2 for details of the reporting parameters and scoring rubric.

Field	Values	Scoring Criteria
Severe Hemorrhage[1]	[Present, Absent]	2 if match ground truth (GT) 0 otherwise
Respiratory Distress[1]	[Present, Absent]	2 if match GT 0 otherwise
Heart Rate[2]	Integer	1 if within *n* of GT 0 otherwise
Respiratory Rate[2]	Integer	1 if within *m* of GT 0 otherwise
Trauma	**Head:** [Wound, Normal] **Torso:** [Wound, Normal] **Upper Ext.:** [Wound, Amputation, Normal] **Lower Ext.:** [Wound, Amputation, Normal]	2 if all match GT 1 if at least two match GT 0 otherwise
Alertness	**Ocular:** [Open, Closed, Not Testable (NT)] **Verbal:** [Normal, Abnormal, Absent, NT] **Motor:** [Normal, Abnormal, Absent, NT]	2 if all match GT 1 if at least two match GT 0 otherwise

Table 6 Preliminary casualty report clinical assessment with scoring criteria

[1] Response receives +2 bonus points if correctly reported within "golden window".

[2] Vitals responses receive +1 bonus point if both are correctly reported within "golden window".

Figure 2: Scoring Rubric

Our robotic fleet (see Figure 1 (a)) included 4 UGVs and 1 UAV - 2 RC cars, one SPOT quadruped, one adapted wheelchair robot, and an Indago 4 drone from Lockheed Martin. The sensor payload, which was consistent across all UGVs, is detailed in Figure 1 (b) - the primary computer being a 64GB NVIDIA Jetson Orin which is heavily optimized for on-edge inference for robotic applications, processing data from disparate sensors (RGB camera on iPhone, radar, microphone, etc.) and publishing predictions to the data network. For all the algorithm developers, this was a key constraint - the model of choice had to be deployable on the arm64 architecture based NVIDIA Jetson Orin. The developed models also had to be packaged into a ros2 node, built as a package, and integrated into the full system such that the executive module could trigger them as needed. Fig. 1 (c) highlights the behavior tree, which is the aforementioned high-level executive module that controls the sequence of algorithm activation during the inspection of a casualty - it ensures that the operator has a full view of which inspection status the robot is currently in, and indicates clearly important markers that are to be noted by the operator before inspection can begin (for example, the 'new casualty ID detected' marker). Robust communications is established between the robots and their respective operator stations using a mesh of Doodle labs radios - this is key for reliable tele-operation, and for reliable reception of predictions at the basestation end. Given the complexity of the described system, a considerable part of our time and energy was spent on field tests (see Figure 5), which allowed us to stress test our communication protocols, models' ability to handle edge cases & real world environments, and that of the operators to perform planned casualty inspections under pressure. In this manner, I contributed to the development, end-end deployment, and testing for the hemorrhage detection and heart rate detection algorithms. The model deployed for the former will be detailed in this thesis; for the latter, a ros2 deployment of ContrastPhys [25] performed the best in all field tests.

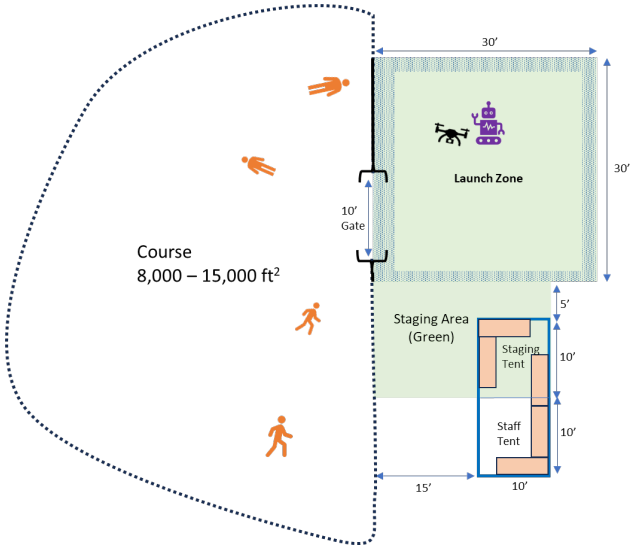

Figure 3: DARPA Triage Challenge Layout Schematic

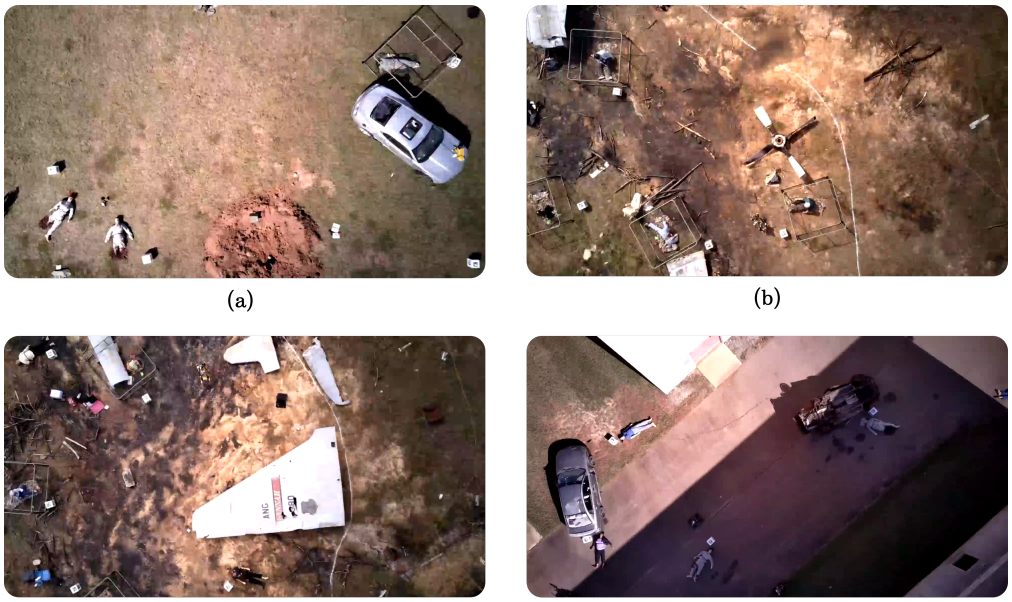

Figure 4: Bird eye's view of sections of the (a) battlefield, (b) & (c) plane crash, & (d) convoy blast DARPA simulated courses.

(a) (b)

Figure 5: Field test snapshots from (a) a demo with Lockheed Martin at Aberdeen, MD and (b) field testing with Pittsburgh EMS at Old VA Hospital, Pittsburgh

Severe hemorrhage was only one of the many health parameters that our robotic fleet was tasked with predicting in the DARPA Triage Year #1 challenge. Other DTC health parameters (not described in this thesis) include alertness parameters (ocular alertness, verbal alertness), presence of amputations, injuries, respiratory distress, and a quantification of vitals (heart rate, respiration rate). Given their importance in predicting need for immediate care, DTC designated severe hemorrhage and respiratory distress as "critical" category parameters, earning 2 points each if predicted correctly. Though my effort over the span of the project included heart rate and respiration rate estimation modules as well, the bulk of my unique effort was focused in severe hemorrhage prediction, and thus this thesis will focus on the same.

4 Methods

Severe hemorrhage, as defined by DARPA for the DTC, involves scenarios with ">50% of the body covered in blood, and/or squirting, bleeding, pooling external to the body." To accomplish this task, we introduce a cascaded framework, **HemoSegFormer**, designed to handle the complexities of hemorrhage detection by first performing segmentation in the image(s) and subsequently classifying hemorrhage severity. Our approach integrates a Feature Pyramid Network (FPN) for blood segmentation and a Mask2Former module for person segmentation, enabling differentiation between blood regions on the body versus external pools or splatters. This implementation was inspired by my prior work with ultrasound images wherein the FPN model was seen to perform well. This section will also detail the other models with which the proposed model will be compared.

In the following discussion, let $\mathcal{I} = \{I_1, I_2, \ldots, I_N\}$ denote a set of N medical images, each represented as an RGB image in $\mathbb{R}^{H \times W \times C}$, with H, W, and C representing the height, width, and number of channels, respectively. The objective of hemorrhage detection and segmentation is to learn a function $f_\theta : \mathbb{R}^{H \times W \times C} \to \{0, 1\}$, parameterized by θ, that maps each image I_i to a binary output indicating the presence ($f_\theta(I_i) = 1$) or absence ($f_\theta(I_i) = 0$) of severe hemorrhage.

4.1 Datasets

The details of the datasets employed in training and testing the models are as below:

- **"Weiss-F8" dataset**: The training dataset consisted of 91 images, collected during field tests wherein fake blood was used to simulate hemorrhage scenarios using mannequins from UPMC and human volunteers from Team Chiron. A majority of the data collected from diverse mannequins was by Dr. Lenny Weiss, which was subsequently annotated by Figure8 Federal - leading us to refer to this dataset as the "*Weiss-F8 dataset*".
- **"DARPA"** or **"Competition" dataset**: During the DARPA competition, we engaged with three distinct simulated scenarios—battlefield, plane crash, and convoy (IED blast). Our initial corpus of data (220 images) composing this dataset was gathered by deploying our

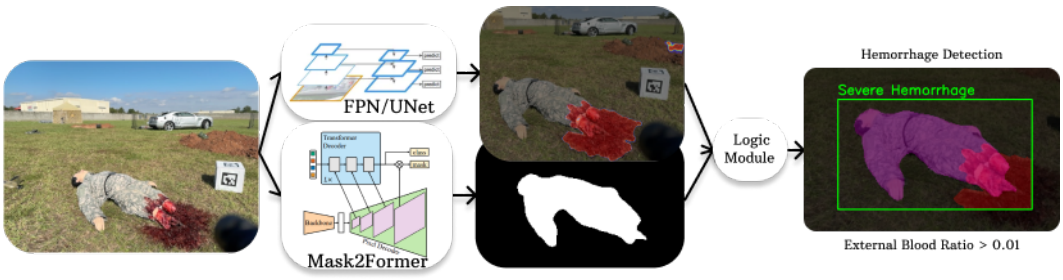

Figure 6: An illustration of the hemorrhage detection pipeline (from left to right) - (i) raw image, (ii) image with predicted blood mask overlaid, (iii) blood mask zeroing applied, person mask overlaid, and hemorrhage presence computed.

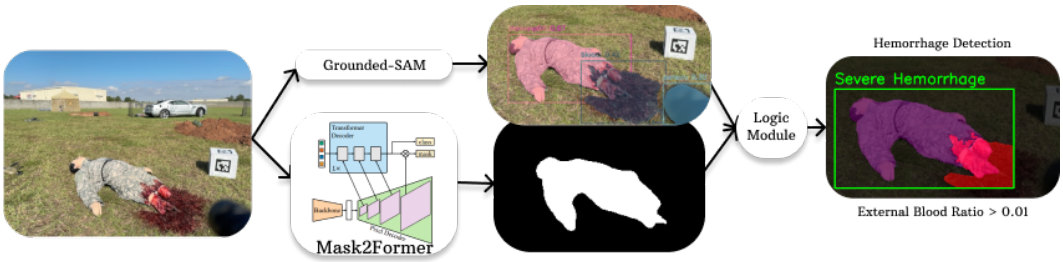

Figure 7: An illustration of the hemorrhage detection pipeline (using Grounded SAM) - (i) raw image, (ii) image with predicted blood mask overlaid, (iii) blood mask zeroing applied, person mask overlaid, and hemorrhage presence computed.

ros2 model node to automatically pick out images from the competition data using the logged ros2 mcap file, to compile a list of the exact same images that the robot used for inference during the competition. While this initial corpus provided a strong foundation, our focus on evaluating model accuracy required minimizing human error in obtaining good viewpoints of the casualties as seen by the robot's sensors. To address this, we removed some images that had obviously bad viewpoints not containing the casualty; then for casualties with few/no good images remaining, we augmented the dataset by manually selecting additional salient frames from the MCAP files that were not processed by the model during competition runs. Through this process, the evaluation dataset was expanded to 249 images, providing a rich representation of the various casualties captured during the competition and thus serving as a fair basis for model evaluation. We refer to this dataset of 249 images interchangeably as both the *"DARPA" dataset* and as the *"competition" dataset*.

- **"DARPA Workshop Eval"** dataset: To make a decision about which model to deploy for the competition, a 'held out' evaluation dataset which was not used in training the models had to be built. This could be built using images from the field tests, but due to a strong risk of overfitting model choice to the field test environment, a set of 20 images - which, collected during a DARPA workshop in June was highly representative of competition conditions - was organized into a small dataset, named "DARPA Workshop Eval". The reason this was not used for training was that the workshop data only represented one of the many possible simulated scenarios; training on this data and testing on the field test data would likely lead to a highly overfit model.

4.2 HemoSegFormer Model Architecture

In our work here, we propose HemoSegFormer (see Figure 6), a cascaded approach to detecting hemorrhage:

- **Blood segmentation:** A Feature Pyramid Network (FPN) model [10] with a ResNet34 encoder [26], pretrained to leverage hierarchical image features, is used for blood segmentation.

This ***hemoFPN*** model captures both fine-grained and high-level contextual information necessary for accurate blood detection. The segmentation function $g_\phi : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^{H\times W}$, parameterized by ϕ, maps an input image I_i to a binary blood mask B_i:

$$B_i = g_\phi(I_i)$$

Note that in other candidate pipelines, this blood segmentation module is the only one that is replaced with the U-Net or Grounded-SAM (GSAM) for blood mask prediction; the other downstream modules as described below are kept intact.

- **Person segmentation:** For person segmentation, we adopt Mask2Former [27], a transformer-based model trained on the COCO dataset. This model is robust in identifying persons in diverse environments, facilitating the differentiation of blood on the body from external blood pools. The segmentation function $h_\psi : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^{H\times W}$, parameterized by ψ, maps an image I_i to a binary person mask P_i:

$$P_i = h_\psi(I_i)$$

  The transformer architecture captures intricate dependencies between image regions, typically enabling accurate person segmentation.

- **Classification (logic module):** A processing pipeline combines the blood mask B_i and person mask P_i to classify the image as "True" or "False" for severe hemorrhage presence. The logic module includes the following steps:
  - **Cropping mechanism:** A bounding box is computed from the detected person mask P_i, and is extended by a factor of 1.2 in both dimensions. Any blood pixels outside this extended bounding box are clipped to zero. Mathematically, the cropped blood mask B_i^{cropped} is defined as:

$$B_i^{\text{cropped}}(x,y) = \begin{cases} B_i(x,y), & \text{if } (x,y) \in \text{ExtendedBoundingBox}(P_i) \\ 0, & \text{otherwise} \end{cases}$$

    This mechanism serves to:
    * Eliminate false positives caused by blood detected on distant casualties.
    * Exclude irrelevant blood regions that do not contribute to the hemorrhage assessment of the current casualty.

    See Figure 8 for a detailed illustration of its utility.
  - **Blood ratio computations:**
    * **Blood external ratio:** Measures the proportion of blood pixels external to the person mask, calculated as:

$$\text{Blood External Ratio} = \frac{\sum_{(x,y)\in\{B_i^{\text{cropped}} - P_i\}} B_i^{\text{cropped}}(x,y)}{\sum_{(x,y)\in P_i} P_i(x,y)}$$

    * **Blood overlap ratio:** Calculates the proportion of blood pixels overlapping the person mask:

$$\text{Blood Overlap Ratio} = \frac{\sum_{(x,y)\in\{B_i^{\text{cropped}} \cap P_i\}} B_i^{\text{cropped}}(x,y)}{\sum_{(x,y)\in P_i} P_i(x,y)}$$

      Classification is determined as "True" if either the blood external ratio exceeds 0.05 or the blood overlap ratio is greater than 0.5. The former condition avoids false positives from slight over-segmentation of blood beyond the person mask; the latter follows DARPA's DTC definition for severe hemorrhage.

4.2.1 Multiview Mechanism for Final Classification and Confidence Scoring

On the robot, a multi-view inference mechanism is implemented, so as to ensure robust final classification. For each detected casualty, the system initiates an initial "monoview" mode, performing a single inference using HemoSegFormer to generate an initial prediction. Following this, a "multiview" mode activates during a hardcoded 30-second casualty inspection, during which the robot operator attempts to drive the robot around the casualty to capture a 360-degree view of the casualty.

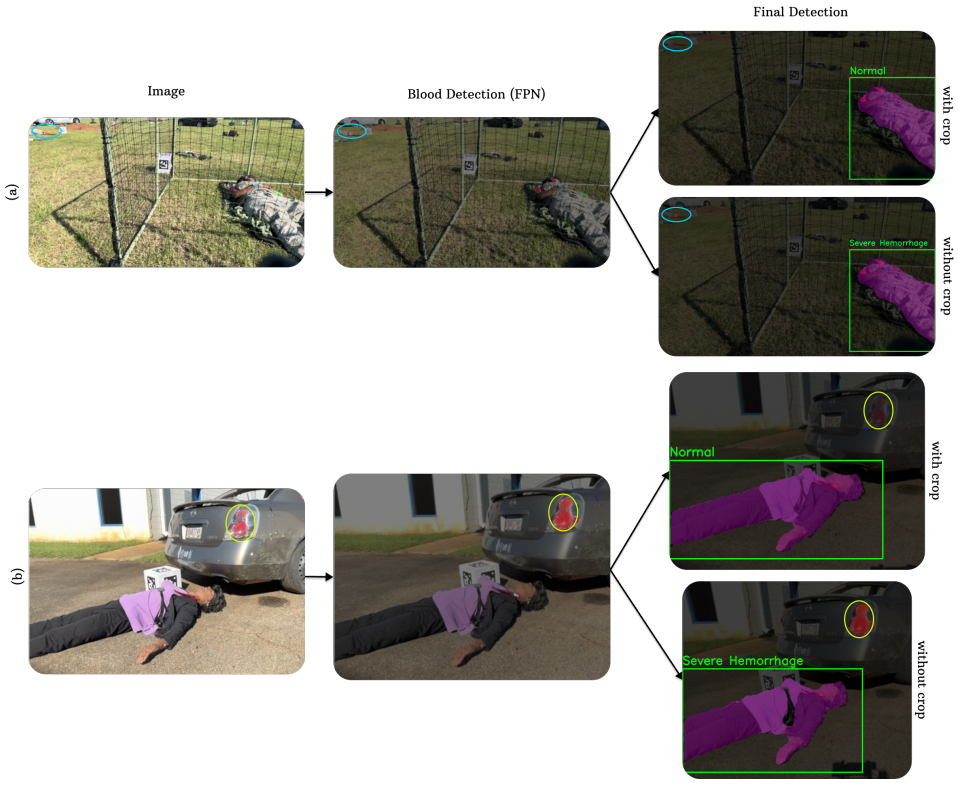

Figure 8: An illustration of the utility of the proximal blood cropping module - (a) a blood pool is detected in a distant casualty with an amputation (blue ellipse, top left); whereas the casualty under assessment has wounds but no severe hemorrhage. The cropping mechanism eliminates distal predictions, leading to a true negative & (b) the hemoFPN model erroneously segments the red brakelights as blood (yellow circle, top right); which is rectified by the cropping mechanism. In both cases it is seen that without the cropping mechanism, a bad prediction is made

In "multiview" mode, the system performs an inference on a single frame every 3 seconds, allowing for a total of 7-10 inferences (based on whether the casualty was kept in view consistently) within the inspection window. To determine the confidence of a positive prediction across these inferences, a confidence score is computed based on the consistency of blood segmentation results. Specifically, given the blood external ratios and overlap ratios from the ~ 7 inferences, the confidence score is calculated as follows:

$$\text{confidence\_score} = 1.0 - \frac{\sigma_{\text{blood\_ext}} + \sigma_{\text{overlap}}}{2}$$

where $\sigma_{\text{blood\_ext}}$ and σ_{overlap} represent the standard deviations of the blood external and overlap ratios across the said inferences. This score provides a measure of prediction consistency, with lower variance indicating higher confidence. Because severe hemorrhage may be diagnosed from blood pooling that is only visible from one of the viewpoints, *if any of the "multiview" inferences predicts severe hemorrhage, the final prediction is output as severe hemorrhage*, and this confidence measure is reported alongside.

4.3 Comparative Models

For evaluating the effectiveness of the HemoSegFormer pipeline deployed in the competition, two additional pipelines were set up using the following models - (i) U-Net and (ii) GSAM. Importantly, note that all of these pipelines used the Mask2Former model for person segmentation and generally followed the above HemoSegFormer pipeline except for replacing FPN for blood segmentation. However, to better understand the performance of the individual blood segmentation models, the evaluation is also done with two versions of each model described: with the proximal blood cropping

module as described in Sec. 4.2 (i) enabled and (ii) disabled. These models were chosen to provide meaningful comparisons, with distinct strengths and limitations highlighted below:

- **HemoFPN:** This is the model deployed in the competition as the blood segmentation module (step #1 as shown in 4.2), leveraging a Feature Pyramid Network (FPN).
  - The FPN's multi-scale feature detection capabilities provide robustness to false negatives.
  - FPN's relatively small model size (26M parameters allows it to learn from fewer training examples.

  With the proximal blood cropping module activated and deactivated, this model pipeline is referred to as **"HemoSegFormer - FPN (w/crop)"** and **"HemoSegFormer - FPN (w/o crop)"**, respectively.
- **U-Net:** A classical segmentation model widely used in medical image segmentation literature.
  - Known for its simplicity and strong performance on medical segmentation tasks.
  - U-Net serves as a proven baseline to contrast with HemoSegFormer, possibly allowing for a highlight of the advantage of multi-scale feature representation in FPN.

  With the proximal blood cropping module activated and deactivated, this model pipeline is referred to as **"HemoSegFormer - U-Net (w/crop)"** and **"HemoSegFormer - U-Net (w/o crop)"**, respectively.
- **Grounded SAM (GSAM):** This novel model combines Grounding DINO's [19] zero-shot detection capabilities with the Segment Anything Model (SAM) for segmentation prompted by language queries.
  - The open-vocabulary detection abilities of Grounding DINO allows it to avoid false detections; by cascading with the strong segmentation ability of SAM, GSAM becomes a strong candidate.
  - **Limitations:**
    - Blood is unlikely to be well-represented in its training distribution, leading to poor generalization for this task.
    - Fine-tuning on hemorrhage-specific data was infeasible within the given time-frame, given the large computational and dataset costs of fine tuning such a large model.
  - **Additional Artifact Removal Heuristic:** During evaluation, it was observed that GSAM often predicted the entire person as the "blood" class. To address this:
    - An artifact removal heuristic was implemented to discard detections from GSAM that overlapped with the person mask predicted by Mask2Former by more than 90% (because it is very unlikely that a casualty would actually be covered with blood on > 90% of their visible surface area).
    - Mathematically, the valid blood mask B_i^{valid} for GSAM is computed as:

$$B_i^{\text{valid}}(x,y) = \begin{cases} B_i(x,y), & \text{if } \frac{\sum_{(x,y)\in(B_i\cap P_i)} B_i(x,y)}{\sum_{(x,y)\in P_i} P_i(x,y)} \leq 0.9 \\ 0, & \text{otherwise} \end{cases}$$

      where B_i is the original blood mask predicted by GSAM and P_i is the person mask from Mask2Former.

    With the proximal blood cropping module activated and deactivated, this model pipeline is referred to as **"GSAM (w/crop)"** and **"GSAM (w/o crop)"**, respectively.
- **DINOv2** (not analyzed in this thesis): DINO v2 [28] is a large transformer-based backbone, trained in a self-supervised manner on extensive datasets so as to learn rich representations of a vast variety of visual data and concepts.
  - Has the ability to capture almost any visual pattern, making it a strong candidate for segmentation in an open world setting.
  - The limitation is that it is not specifically tuned for medical segmentation tasks like hemorrhage detection. More details of our effort in finetuning and testing DINOv2 [28] that led to its exclusion from the final evaluation are captured in the Section 4.3.1.

The performance of these pipelines is compared to the HemoSegFormer cascade to assess their ability to accurately detect severe hemorrhage. The proximal cropping-disabled versions serve as benchmarks to evaluate the pipelines' raw segmentation ability and their susceptibility to false positives without additional logic. This leads to a total of six pipelines being tested on the evaluation dataset, allowing us to study across the axes of model choice and engineering choices.

4.3.1 DINO V2

Given its training on a vast corpus of visual data, DINOv2 is ubiquitous in CV applications, wherein it has been employed in various vision tasks like zero-shot detection of geographical location from blurry drone images[29], zero-shot object-level image customization[30], and as the basis for development of a foundational models for cancer detection[31]. We finetuned it on the "Weiss-F8 dataset"(see Sec. 4.1), and then tested it on a small held out dataset - samples from which are presented here to justify its early removal from the model candidate pool.

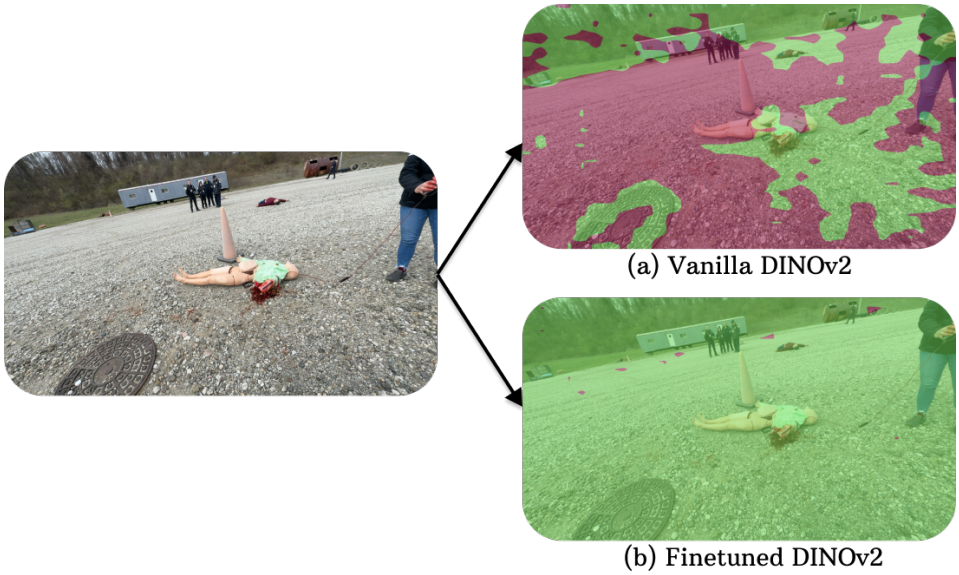

Figure 9: Inference results using a eval data sample on - (a) an untrained DINOv2 model, and (b) DINOv2 finetuned on Weiss-F8 data.

As seen in Figure 9, the first inference was done on a untrained DINOv2 model as a sanity check - initialized with weights from the "facebook/dinov2-base" checkpoint and with a linear classifier for pixel-wise predictions. Though the decoder is only a randomly initialized simple linear layer (nn.Linear), the richness of visual features in the large transformer backbone is evident, as the model weakly clusters the image based on patterns (the fake blood line running from the mannequin to the top right of the frame is roughly followed by a segmentation mask, for example). Next, we fine-tuned the model for 100 epochs using the AdamW optimizer with a learning rate of 0.005 and cross-entropy loss, ignoring background pixels. During training, the DINOv2 backbone was frozen to preserve its pre-trained feature extraction capabilities while fine-tuning the segmentation head. Despite these efforts, as seen in (b), the model's performance on semantic segmentation was poor, with no clear segmentation of blood or even blood-like entities. From this, and from literature on finetuning decoders for DINOv2, it is clear that a much larger data corpus is needed to be able to use the DINOv2 model for our task. Since that was untenable, this model was quickly eliminated as a possible candidate.

4.4 Training Procedure and Loss Function

We train the hemoFPN model using a DICE loss function, a popular choice for segmentation tasks that balances between foreground-background class imbalance and emphasizes overlap between predicted and ground truth masks:

$$\mathcal{L}_{\text{DICE}}(I, B) = 1 - \frac{2\sum_{(x,y)} B_i(x,y)I(x,y)}{\sum_{(x,y)} B_i(x,y)^2 + \sum_{(x,y)} I(x,y)^2}$$

where B_i denotes the predicted blood mask, and I is the ground truth mask. DICE loss encourages a high degree of overlap in the segmented regions, minimizing false negatives in cases of small or faint blood regions.

Training was conducted for 100 epochs using 5-fold cross-validation to evaluate model generalization and mitigate the risk of overfitting, particularly given the limited dataset size. In each of the folds, validation loss was tracked to pick the best model. An AdamW optimizer was employed, paired with a learning rate scheduler to dynamically adjust the learning rate based on the validation loss, facilitating efficient convergence. The "Weiss-F8 dataset", as detailed in Sec 4.1 was used as the training dataset. To enhance generalization and address the limited dataset size, data augmentation techniques were applied, including random translations, rotations, and brightness adjustments. The initial learning rate was set to 0.001, and the learning rate scheduler (ReduceLROnPlateau) was set to a configuration of patience parameter of 5 epochs and a reduction factor of 0.5. The GSAM model used the 'base' checkpoints for Grounding DINO 1.5, and Segment Anything Model (SAM), as provided by Huggingface, with no further fine-tuning.

4.5 Evaluation Metrics

We evaluate the HemoSegFormer performance based on two primary metrics:

1. **DARPA Points**: DARPA defines a binary ground truth for severe hemorrhage for each casualty. For each correct model prediction, the team earns 2 points, while incorrect predictions yield 0 points. A total DARPA score is computed for each of the three DARPA simulated runs (battlefield, plane crash, convoy IED) for each model. This metric provides a direct measure of the model's effectiveness in real-time triage scenarios.

2. **DICE Score**: Based on ground truth segmentation masks provided by Figure 8, DICE scores are computed to evaluate segmentation quality. For each run, mean DICE scores across the dataset are calculated, allowing for a comprehensive analysis of segmentation accuracy.

While DARPA points provide a task-specific binary evaluation, the DICE score allows for quantitative assessment of segmentation quality at a pixel level. Together, these metrics offer a holistic view of both model robustness in real-world settings and segmentation precision.

4.5.1 Evaluation Dataset

To rigorously evaluate the models, it was essential to construct a dataset that fully encapsulates the task's inherent challenges. We performed our intial evaluations using our "DARPA Workshop Eval" dataset, a small held-out dataset collected during a DARPA workshop in June, informing our selection of HemoSegFormer (FPN) as the model of choice for deployment in the competition. Subsequently, after attending the DARPA competition in October, we were able to accumulate a much larger corpus of data, which was processed as described in Section 4.1 to form our final held-out "DARPA dataset". All candidate models/pipelines were evaluated in this present work based on their performance in this "DARPA dataset".

5 Results

As discussed earlier, the candidate pipelines described in 4.3 were all evaluated on the evaluation dataset gathered from the DARPA Challenge Event #1, as described in 4.5.1. The evaluation was done on a P-5000 GPU, with scripted python calls, so as to enable higher iteration speed. Although not presented in detail here, the performance of these AI models was also verified in an equivalent pipeline that replicates true deployment of the AI models - this pipeline replicates the full robotic

system's AI software stack, including the ros2 nodes, to which it fed as input the ros2bag recordings from the robot runs at the DTC event.

Firstly, preliminary tests conducted on the "DARPA Workshop Eval" dataset revealed the following expected model behaviors:

- **HemoFPN and HemoU-Net:**
  - These models reliably detect blood patches, even from a distance, ensuring highly sensitive hemorrhage detection.
  - A known limitation is their tendency to misclassify 'blood-like' entities (e.g., red clothing, flags) as blood, leading to false positives.
- **GSAM:**
  - This model demonstrated the ability to detect blood pools but with inconsistent results.
  - In some cases, blood pools were missed in one frame but detected when the viewpoint shifted - an inherent variability that stems from the likelihood that detecting blood is a out-of-distribution task for this model.

Table 1: Model Performance Comparison on DTC Official Simulated Courses

Course	Model - (see Sec. 4.3)	Casualties	Correct Predictions	Accuracy (%)
Battlefield	**HemoSegFormer - FPN (w/crop)**	12	**9**	**75.0**
	HemoSegFormer - U-Net (w/ crop)		**9**	**75.0**
	HemoSegFormer - FPN (w/o crop)		5	41.7
	HemoSegFormer - U-Net (w/o crop)		6	50.0
	GSAM (w/ crop)		7	58.3
	GSAM (w/o crop)		6	50.0
Plane Crash	HemoSegFormer - FPN (w/crop)	8	3	37.5
	HemoSegFormer - U-Net (w/ crop)		3	37.5
	HemoSegFormer - FPN (w/o crop)		3	37.5
	HemoSegFormer - U-Net (w/o crop)		3	37.5
	GSAM (w/ crop)		**4**	**50.0**
	GSAM (w/o crop)		3	37.5
Convoy	**HemoSegFormer - FPN (w/crop)**	7	**7**	**100.0**
	HemoSegFormer - U-Net (w/ crop)		5	71.4
	HemoSegFormer - FPN (w/o crop)		4	57.1
	HemoSegFormer - U-Net (w/o crop)		3	42.9
	GSAM (w/ crop)		5	71.4
	GSAM (w/o crop)		3	42.9

The results from evaluation on the full "Competition dataset" (as gathered from the DTC official simulated courses, detailed in 4.5.1) are presented in Table 1.

The HemoSegFormer (FPN) model with cropping enabled, performed the best on two out of three of DTC's simulated courses - achieving 9 casualties correct of the 12 inspected in the battlefield run (75% accuracy) and achieving 7 correct out of 7 inspected in the convoy run (100% accuracy). The plane crash course was the most difficult for all of the models tested due to the fact that (i) many casualties were human volunteers wearing red clothing and/or holding red flags (see discussion in 6 for a deeper treatment), and (ii) input data was of lower quality due to navigation difficulties. Although GSAM with cropping performed the best on the plane crash run, the HemoSegFormer achieved only 1 fewer correct predictions. Overall, the HemoSegFormer (FPN) model earned the most DARPA points of any model tested across all 3 runs combined, and it also earned the best overall DICE score of 0.419, as seen in Table 2.

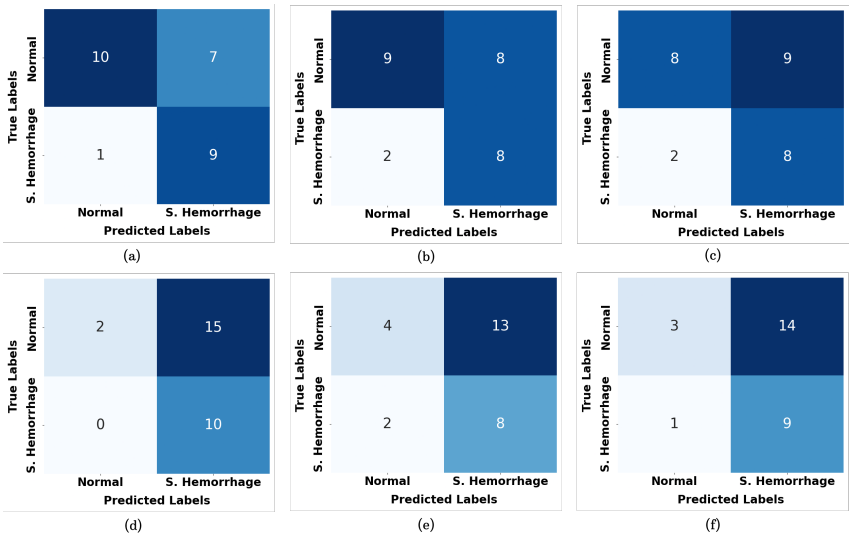

Figure 10: Confusion matrices derived from evaluation on the full competion dataset using - (a) HemoSegFormer - FPN (w/ crop), (b) HemoSegFormer - U-Net (w/ crop), (c) GSAM (w/ crop), (d) HemoSegFormer - FPN (w/o crop), (e) HemoSegFormer - U-Net (w/o crop), (f) GSAM (w/o crop).

As for the U-Net model - it was noted that it performed slightly worse than the FPN, as reflected in its overall DICE scores of 0.414 and 0.353, as compared to the scores of 0.419 and 0.359 achieved by the hemoFPN- with the blood cropping module activated and deactivated, respectively. As for competition scoring, however, which was a less granular but more important metric - it was on par with hemoFPN on the battlefield run, and plane crash run, achieving 9 casualties correct of the 12 inspected, and 3 correct out of 8 inspected, respectively. In the convoy run, however, it achieved 5 correct predictions of 7 casualties inspected, as compared to the 7 on 7 correctness achieved by the hemoFPN. These failure cases are discussed in Section 6.

As noted earlier, the GSAM model was not expected to perform well, given that it was not observed to pick up blood pools reliably. However, it is surprising to note the success rate it had on the larger competition dataset - it achieved 7 of 9 casualties correct on the battlefield course, 4 of 8 casualties correct on the plane crash course, and 5 of 7 casualties correct on the convoy course - only two predictions off in both respective runs from the top-performing HemoSegFormer (FPN). It was quickly discovered that despite its erratic ability to segment blood, the multiview inference logic (described in Sec. 4.2.1) created an ensembling affect, and in cases of severe hemorrhage the model often picked up the blood mask correctly in at least one of the attempts. A few non-trivial positive cases will be presented in Section 6.

Table 2: Overall Blood DICE Scores, Course-wise Blood DICE Scores & Overall DARPA Scores

Model	**Overall Blood DICE**	**Blood DICE Score**			**Overall Score**
		Battlefield	**Plane Crash**	**Convoy**	
HemoSegFormer - hemoFPN (w/crop)	**0.419**	**0.594**	0.168	**0.471**	**38**
HemoSegFormer - U-Net (w/ crop)	0.414	**0.577**	0.162	**0.476**	34
HemoSegFormer - hemoFPN (w/o crop)	0.359	0.441	0.155	0.335	24
HemoSegFormer - U-Net (w/o crop)	0.353	0.546	0.125	0.311	24
GSAM (with crop)	0.368	0.361	**0.646**	0.035	32
GSAM (no crop)	0.355	0.388	**0.674**	0.018	24

Table 2 presents the DICE scores computed in a run-wise manner - i.e., the DICE score is computed for the blood mask prediction on each viewpoint of each casualty, and averaged across all the viewpoints of all casualties for a given run. This lets us gain insights into the performance of models across runs without going into the granularities of per-person DICE. It is notable that the DICE scores recorded in Table 2 act as a strong predictor of the DARPA competition performance indicated in

Table 1. However, a few interesting outliers do exist - in the plane crash course, the GSAM model (no crop) has a higher DICE (0.674) than the GSAM model with cropping activated (0.646); however it scored lower (3/8 casualties correct) than the said GSAM (crop) model (4/8 casualties correct). This points to the implication that the cropping mechanism is likely eliminating false positives; however, the though the higher DICE score of the 'no crop' pipeline can be explained by the correctly predicted blood mask remaining uncropped (see Fig 15 (c) for reference of how cropping leads to a lower DICE score though the final prediction remains unaffected). This points to an interesting tradeoff - simply put, it was acceptable for the cropping mechanism to lead to some parts of the external blood detected to not play a role in the final 'severe hemorrhage' or 'normal' prediction (which would not be affected at all), as long as it served its original motive described in 4.2.

6 Discussion and Analysis

As previously described, our custom HemoSegFormer (FPN) model was trained on the custom "Weiss-F8" dataset and evaluated in DARPA's Event #1 challenge. HemoSegFormer's overall setup - (i) using a custom FPN for blood segmentation, (ii) cascading with Mask2Former to calculate overlap and perform proximal blood cropping, & (iii) using multiview inference to capture occluded blood pools and provide an ensembling effect to improve inference robustness - led the hemorrhage module to score a grand total of 38 points in the DARPA Year#1 event.

In general, all the correct predictions performed by the hemoFPN, hemoU-Net and their cascaded variants, looked like the illustrations in Figures 6, 15, 17, and 18 - which exemplify HemoSegFormer's consistent ability to pick up blood, whether pooled or on clothes, thus explaining its high task accuracies as see in Table 1.

6.1 A deep dive on model failures

Due in part to the diverse field conditions posed by the DARPA Triage Challenge, there are multiple types of failures for the hemoFPN, hemoU-Net and GSAM models. Since the hemoFPN and U-Net behave similarly, between the two of them this analysis will focus on the failures of the better-performing hemoFPN model. HemoFPN's failures in the battlefield course vs. its failures in the plane crash course point to two uniquely differently modes of failure, as discussed below.

Figure 11: An illustration of a failure case - the right arm is amputated, with a dark pool of blood in the shadows (encircled). The hemoFPN model was unable to pick this up.

Figure 11 presents an inherently challenging case, since the blood pool is very dark and is hidden in the shadows. Not only is the blood pool hard to see, this situation is far from the training distribution, so it is unsurprising that the hemoFPN pipeline incorrectly classifies the casualty as "Normal".

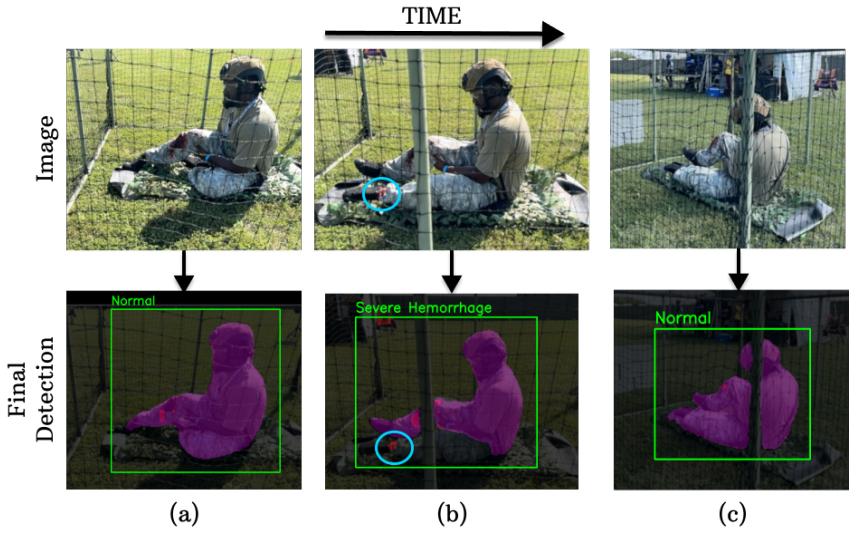

Figure 12: An illustration of a person detection failure case (see encircled area for a true positive blood stain) misinterpreted by logic module as a false positive "severe hemorrhage" due to a partial under-segmentation of the person's leg, which leads the AI to incorrectly treat the blood as pooling external to the body.

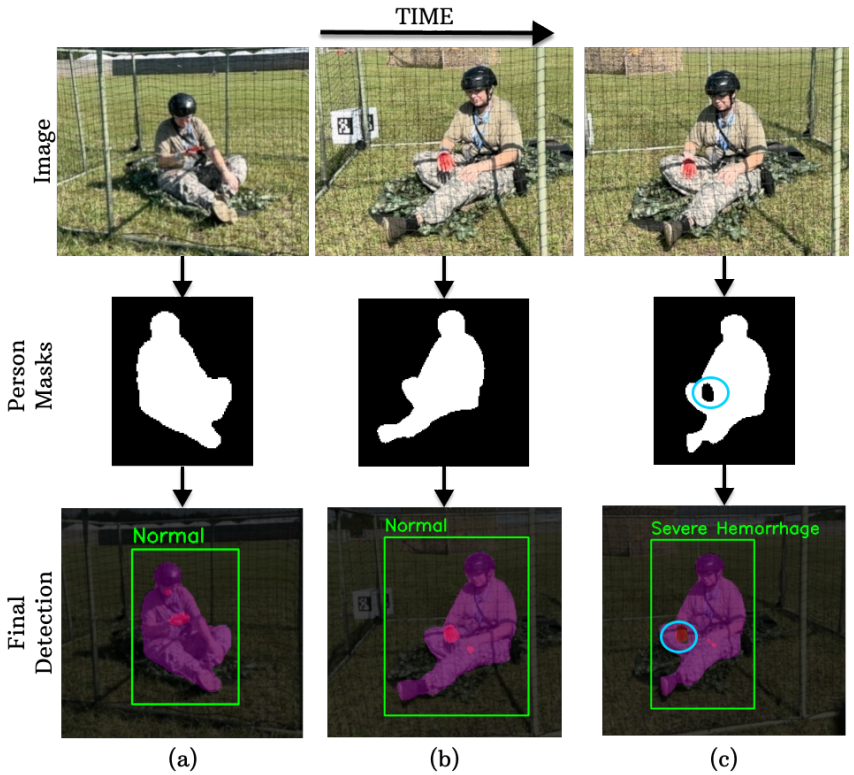

Figure 13: Person under-segmentation failure case #2 - see encircled area in (c) for the person segmentation flaw. In contrast to Figure 12 where the off-the-shelf human-segmentation AI model missed an entire leg, in this case the human segmentation failed only where the blood covered the body.

Figure 12 presents a different type of failure, in which the blood prediction worked perfectly, as shown in (a) and (c), but the overall AI pipeline failed in (b) due to a partial failure of the person detection model not segmenting the entire person; the blood detection is still correct, but the blood stain on the left leg is incorrectly considered by the model as "external to casualty", leading to an incorrect "severe hemorrhage" detection. A similar pattern is seen in the case of Figure 13, where the casualty in the picture has blood on her hands. As seen in (a) and (b), the blood detection model correctly picks up the blood mask, and given that it is on the person, the computed overlap ratio is less than 0.5 - a "Normal" detection. As the robot circles the casualty, looking to catch any occluded pools of blood or evidence of hemorrhage, there is a single failure of the person mask as illustrated in Fig 13 (c); again the blood mask is correct, but the overall AI pipeline incorrectly treats this as an "external blood pool", leading to a an incorrect "severe hemorrhage" classification. In contrast to the previous example where pre-trained Mask2Former's human segmentation missed an entire leg (a fundamental failure of the model), in this case Mask2Former only failed where blood covered the body (an out-of-domain-distribution failure).

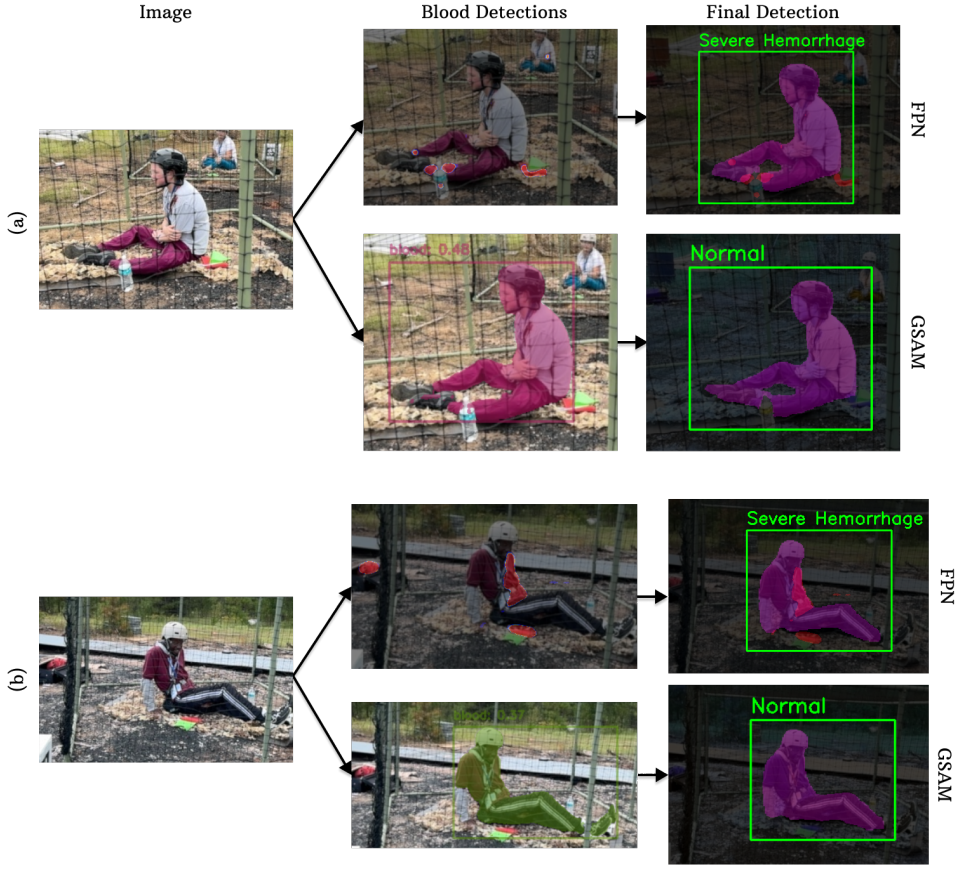

Figure 14: FPN has a tendency for false positives in the presence of objects with "blood-like" appearance. GSAM's learned semantic grounding leads to superior inference accuracy in such cases.

The plane crash course, on the other hand, contained multiple "blood-like" entities, e.g. red sweaters, flags, etc. Although hemoFPN was otherwise fairly reliable in segmenting blood, its learned features/heuristics appear to be insufficient to differentiate between blood vs. objects with blood-like appearance. GSAM, on the other hand, is a semantically grounded model which can be expected to recognize red clothes as clothing and otherwise recognize a wide variety of blood-colored objects, thus (one would hope) avoiding false-positives or over-segmentation of blood. In some cases, GSAM does indeed perform better than hemoFPN or hemoU-Net. As seen in Figure 14 (a) and (b), shadows in the casualty's maroon colored pants, and a red flag next to them, are incorrectly segmented by the hemoFPN model as blood, leading to a false detection of "severe hemorrhage". In both (a) and (b), GSAM also detects blood; however, it segments the entire person as "blood" with low confidence.

As explained earlier, this is a known issue, and the artifact removal heuristic described in Sec. 4.3 discards the same. There are no further blood detections, and the GSAM-based pipeline correctly classifies the casualties (a) and (b) as "Normal". Alongside avoiding false positives, GSAM is also able to achieve true positives; as shown in Figure 15, the GSAM model has two detections for the "blood" query - one covers the whole person with a low confidence of 0.35, and the other covers the blood pool precisely with a slightly higher confidence. As mentioned before, the artifact removal heuristic checks for this expected behavior by comparing the overlap of individual detections with the Mask2Former person mask and if overlap is greater than 85% (as compared to the person mask) the detection is discarded. If the GSAM model had been able to detect many/most true positives (while continuing to avoid false positives), then it would have achieved the best accuracy by a large margin on the plane crash course. However, as is clear from the plane crash results in Tables 1 and 2, the GSAM model outperformed the HemoFPN and U-Net models by only a small margin, achieving 4 correct predictions out of the 8 human casualties assessed; whereas all other models and their variations achieved 3 correct of the 8 assessed casualties.

All failures of the GSAM model in plane crash course are concurrent with an unexpected failure in the artifact removal heuristic. Figure 16 presents an interesting comparison of the failures of GSAM vs. hemoFPN in 3 challenging cases in the plane crash course. In (a), the hemoFPN model sees a red T-shirt in the distance on a different casualty, detects it as a blood pool external to the proximal casualty body, leading to a "severe hemorrhage" assessment. GSAM has a different failure, detecting the whole person as a blood mask; the artifact removal method fails here due to over-segmentation by the Mask2Former model, leading to the overlap ratio falling just short of 85%, which then leads to that blood detection artifact being retained, resulting in a false "severe hemorrhage" prediction. The same patterns repeat in (b) and (c). In these cases, the hemoFPN model failed consistently across viewpoints, with false positive output in all 8 of its multi-view inferences in each of (a), (b), and (c) - courtesy the background casualty, red flag on the ground, and bright red shirt, respectively. GSAM did better on many viewpoints; the example (c) was the only failure among its 8 viewpoints, with GSAM correctly predicting "Normal" for the other 7 viewpoints (not shown here).

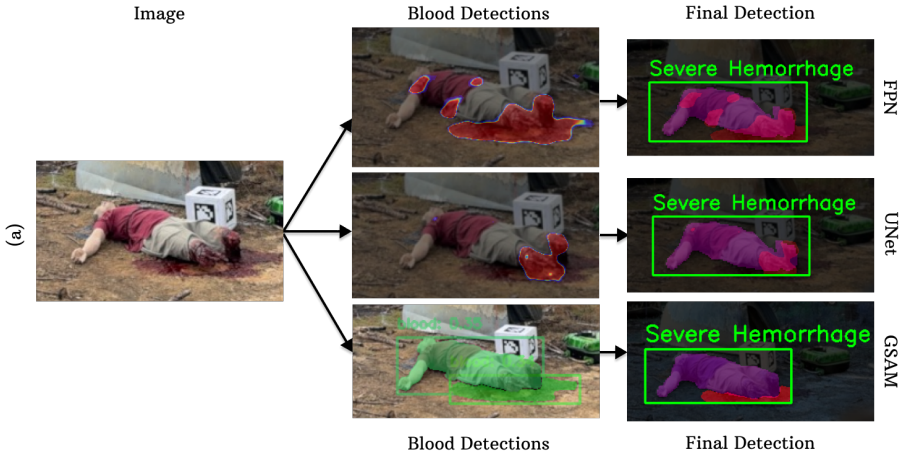

Figure 15: FPN vs U-Net vs GSAM - an illustration of consistent success

In this tradeoff between segmenting all the blood vs. excluding other blood-like objects, there are some noteworthy differences between hemoFPN vs. U-Net. In Figure 15, note that the hemoFPN model segments the blood pool accurately; but also starts to segment parts of the red tshirt on the casualty as blood. The U-Net, however, is more conservative with segmenting the blood pool and only partly achieves the correct blood segmentation (almost ending up with a false "Normal" classification) - however, the same conservatism also seems to let it avoid segmenting the red tshirt as blood. This tradeoff between the hemoFPN and U-Net is seen consistently across the dataset and partially explains their extremely close DICE scores. The same behavior is seen in the example from run #3 - see Figure 17. In that case, however, the conservative U-Net leads to a false "Normal" prediction whereas the hemoFPN model correctly gets the head trauma and blood pool near the casualty's head - a correct "severe hemorrhage" result. This might imply that U-Net is slightly less prone to false positives than the FPN, which is corrroborated by by comparing the hemoFPN and

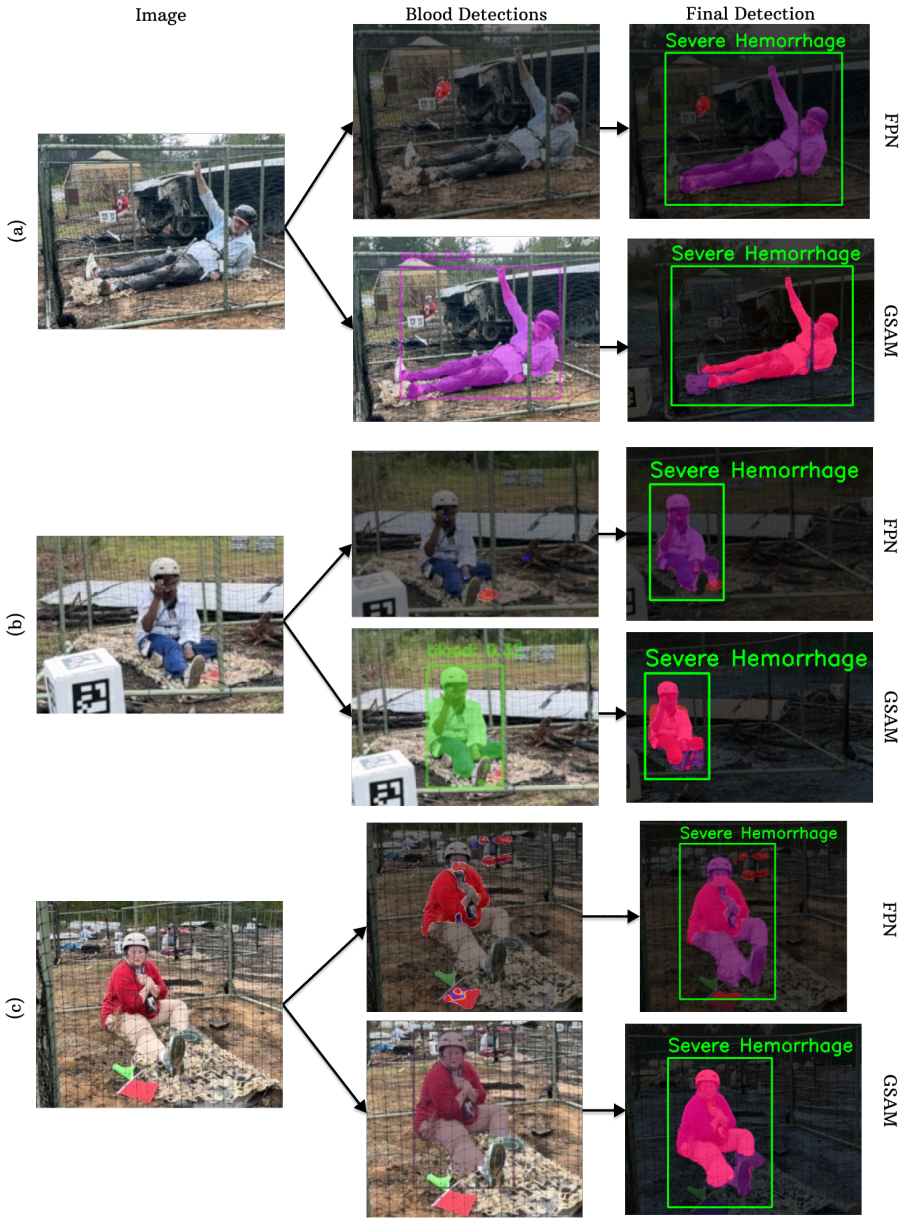

Figure 16: FPN vs GSAM - an illustration of an unexpected artifact removal heuristic failure for GSAM pipeline. HemoFPN exhibits expected (but incorrect) behavior of segmenting red entities as blood.

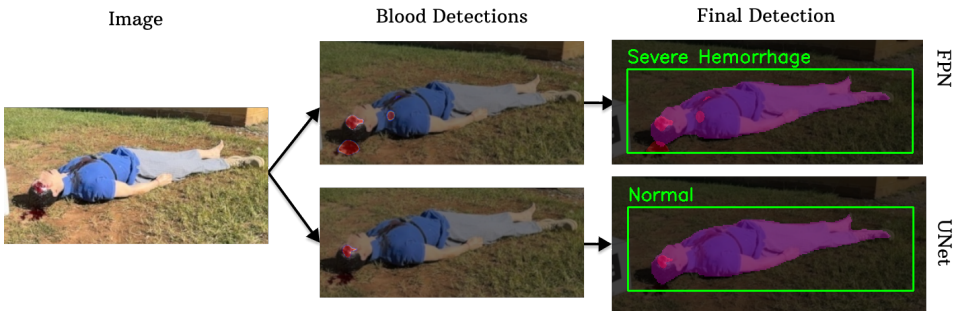

Figure 17: An illustration of the difference in segmentation conservatism levels of U-Net vs hemoFPN models.

U-Net models without involvement of the proximal blood mask cropping (see Figure 10 (d) and (e)). Overall, with the involvement of the blodd cropping module, hemoFPN appears to be superior to U-Net for the hemorrhage detection.

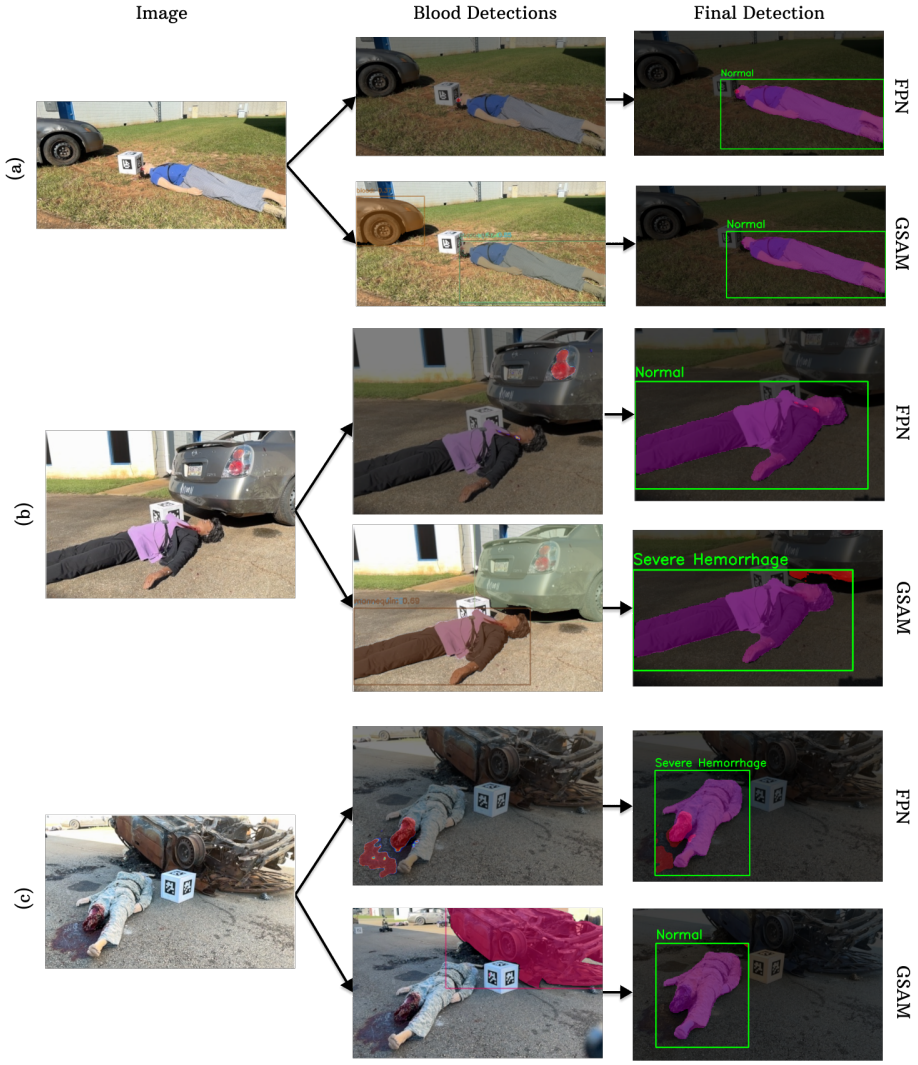

Figure 18: FPN vs GSAM - an illustration disproving the assumed superiority of the out-the-box GSAM's semantically grounded segmentation abilities.

For the convoy course, hemoFPN had a 100% success rate- thus, only GSAM's failures will be discussed for that course. Referring to Figure 18, there is a head wound visible in (a) that hemoFPN detects, whereas GSAM model has two blood detections - one that covers the whole car in the background and one that covers the casualty fully. The artifact removal heuristic discards the latter; and the proximal blood cropping discards the former. Although GSAM has the same output as hemoFPN for this view, GSAM completely failed to correctly segment the blood. It is also worth noting that in other viewpoints (not shown) for this casualty, a blood pool becomes visible, making "severe hemorrhage" the correct classification, and in multiview inference, as shown in Figure 17, the hemoFPN is able to correctly predict the same. In Figure 18 (b) the car's brake lights pose an expected challenge for the hemoFPN model - it captures the rear brakelights of the car as a blood mask alongside correctly capturing the blood on the casualty's neck, but the former is discarded by the blood cropping module, leading to a correct "Normal" prediction. The GSAM model, however, captures the whole car as a blood mask; and though most of it is discarded by the blood cropping, enough of it is retained by the blood crop to erroneously lead to a "severe hemorrhage" detection. In (c), the hemoFPN correctly segments the blood pool external to the casualty, whereas the GSAM model again missed the blood pool; its only blood detection (precisely) covers the rusted car in the background, which is then discarded by the proximal blood cropping module, leading to a false negative "Normal" prediction. Comparing between the versions of GSAM with and without the proximal blood cropping module shows a significant accuracy difference as seen from the Tables 1 and 2. A sizeable portion of GSAM's false positives stem from detections that are far from the focus casualty, as evidenced by Figure 10 and the above discussion, once again showing the utility of the proximal cropping heuristic in reducing false positives.

6.2 The need for grounding

As discussed earlier, although smaller models like the FPN can be trained for unique applications from more limited training data than could a large foundation model, when these smaller models need to be deployed in a real world setting, there is a need for integration with a bigger pre-trained model that has a more "general intelligence", based on its training on vast amounts of image and language data.

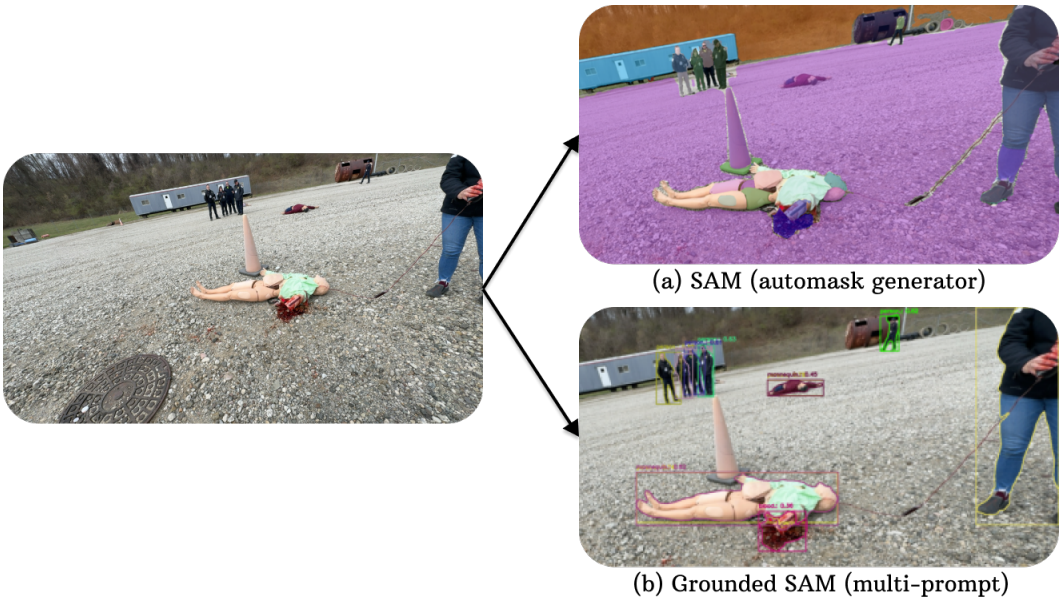

Figure 19: Inference results using a data sample on - (a) SAM with automatic mask generation, and (b) GSAM, with {"person", "mannequin", "blood", "red"} prompt sequence

Figure 19 illustrates some of the "open-vocabulary" detection abilities of SOTA models in this realm. In (a), the SAM model is able to segment multiple different instances, including a good segmentation of the blood. Likewise, in (b) GSAM does a good job segmenting the people, mannequins, and

blood in the frame. These examples illustrate the potential for foundation models to perform open-vocabulary segmentation of a world scene with no finetuning. However, as the previous results have shown, foundation models are imperfect at performing open-vocabulary analysis of images/videos outside the domain of their large corpus of training data, and so it remains an open challenge to better harness their semantically grounded visual representations for unique purposes. Our work highlights this challenge, which we (and the larger AI community) are working to address.

7 Future Work

DARPA's Triage Challenge forced us to assess and advance state-of-the-art technologies for unique and challenging problems. Although it is normal in research to push the "edge" of what is possible, we have been required to push on multiple "edges" at once.

One major limitation encountered with small models (FPN [10] and U-Net [3]) was their tendency to misclassify red-colored objects—such as clothing, flags, and even flushed skin—as blood. An apparent simple solution seemed to be to deploy a semantically grounded model such as Grounding DINO [19] or LLAVA [32]. Since these these models leverage joint language-vision embeddings, similar to how the human brain forms semantically enriched spatial representations [33], it was assumed that they would easily disambiguate between red-colored artifacts and actual blood. However, although GSAM [18], which is one of the most advanced semantically grounded vision models in existence, was successful to a fair extent, it also segmented cars and windows as blood, thus making it unreliable for deployment. Given the size of our original training dataset and timeline constraints, finetuning GSAM or MedSAM was not possible - it remains to be seen if the competition data can be heavily augmented and split into train-test sets, to train and evaluate finetuned versions of said models. On the topic of performance of small models like hemoFPN and U-Net, one other simple axis to explore is that of dataset size - a data ablation study could investigate whether simply increasing the dataset size could imbue the model with the ability to differentiate between blood and red objects. If inconclusive, this would further validate the need for grounded segmentation pipelines.

One attractive solution would be to create a hybrid pipeline which combines the detailed visual feature detection abilities of hemoFPN (or its variants) and the open-vocabulary semantic understanding of GSAM (or equivalent). Given the diverse failures of all the models involved, naive late-fusion heuristics seem insufficient to reconcile the predictions of hemoFPN and GSAM. The question of how to integrate the abilities of GSAM and hemoFPN in an "early fusion" manner remains an open research problem given the inherent differences in their architecture, and it is one I am excited to delve into in the coming months. An emerging consensus among our team is to create and test a novel end-to-end training setup by using a large-encoder backbone rich in visual features (like DINOv2), with the FPN head attached as a randomly initialized decoder, and with skip connections drawn as links from DINOv2's intermediate features. As of now it is unclear how much data will be required to reasonably train such a complex setup, but we will continue data collection this spring and systematically begin the aforementioned data ablation study on this architecture.

The current success of the proximal cropping module demonstrates the potential of heuristic-based enhancements. However, these heuristics are brittle and can fail easily, as seen in the failure cases of the GSAM model - one unique line of future research would be in developing adaptive heuristics trained on a large annotated dataset - this could minimize errors caused by artifacts like elongated person masks or distant objects and improve hemorrhage prediction accuracy. Learnable heuristics seems to be an under-explored direction of work ([34]) in biomedical AI, and given our practical application case it is our hope that exploring it could lead to novel insights for the community. In Year #2, we expect DARPA to present us with dynamic hemorrhage scenarios instead of the static scenarios presented in Year #1, so we aim to incorporate temporal reasoning across multiview inferences, to allow more fine-grained reasoning about changes in the visual features detected in the inferences.

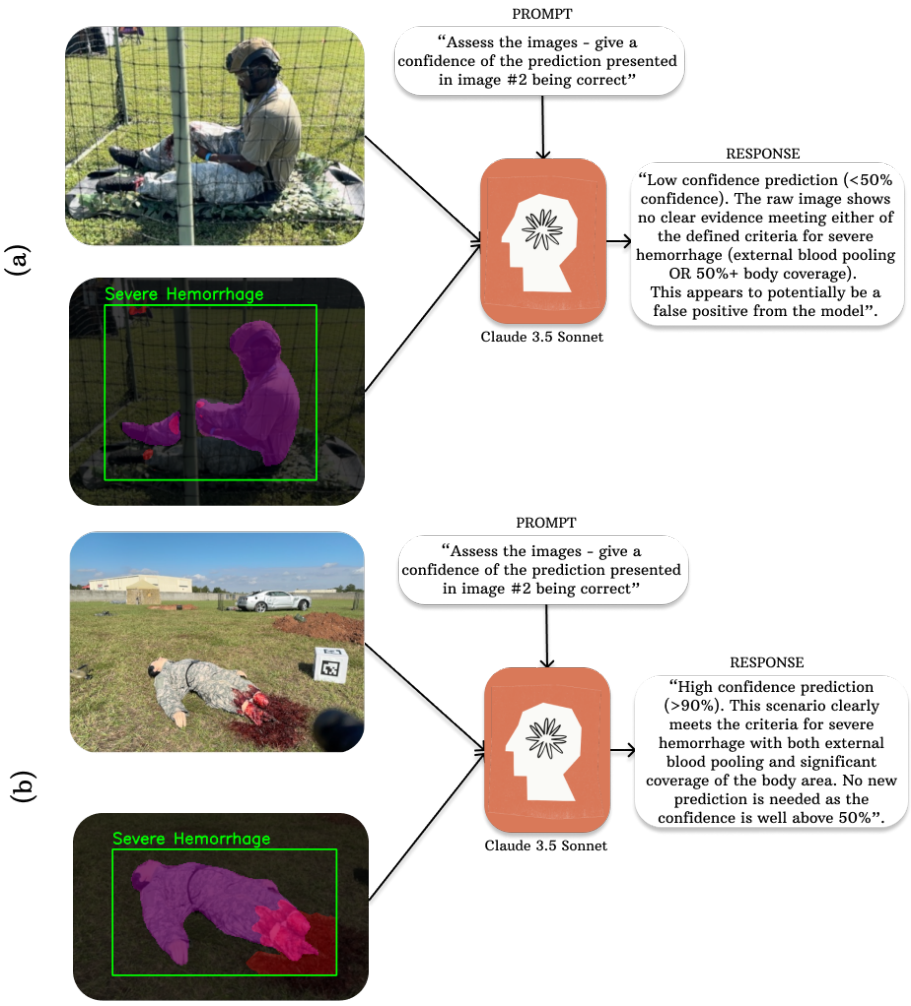

Figure 20: An illustration of Claude's 3.5 Sonnet model being able to reason about hemorrhage predictions and assign confidence.

Another avenue for combining the domain-specific abilities of hemoFPN (or its variants) with the semantic understanding of VLMs such as GSAM is for the VLM to act as an expert-in-the-loop, assigning confidence to the assessments of hemoFPN as shown above in Figure 20. To implement inter-segmentation and input-vs-results reasoning on edge, we have begun looking into the workings of recent models like ViP-LLAVA [35], which are trained on images containing each of 8 different visual prompts (including mask contour, ellipse, bounding box, etc.), such that the model learns to attend to very specific parts of an image, as pointed out by a human user, and then reason about the defined region of interest. Benchmarks, as reported in the ViP-LLAVA [35] work, show that such novel open-source models are performing even better than general-purpose models like GPT-4V [36] at fine-grained video understanding tasks. We plan to integrate such a model into our pipeline to perform pixel-wise reasoning across the predicted segmentation masks, with the predicted masks of one class acting as attention masks for the predicted masks of other classes; we theorize such an arrangement could substantially improve the hemorrhage prediction accuracy. For example, feeding such a model with a trio of inputs consisting of the raw input image, FPN-predicted blood segmentation as primary mask, and the person segmentation as an attention mask, could lead to more intelligent rejection of the false positives and false negatives seen in the preliminary results section above. It will be interesting to explore whether multimodal models like ViP-LLAVA [35] can achieve more human-level reasoning about the relationships between images and segmentations.

Lastly, as we consider inter-image reasoning using a multimodal VLM, it is also interesting to consider the related concept of utilizing semantic information to perform "informed segmentation".

There is little work in this space, presenting a novel avenue for research. (In our prior work developing AI tools for pleural effusion detection, the clinicians were often able to reason in natural language about ambiguous cases, discussing out loud what parts of the image they were looking at to diagnose an effusion versus those image regions that counter-indicate such a diagnosis.) If such rich semantic information could be fed into an AI pipeline performing segmentation and serve as an "attention mask" to improve accuracy, it could represent a technological leap forward in being able to work quickly with challenging cases. For hemorrhage detection, a VLM could be used on the raw input video stream to generate (i) a scene description, (ii) initial segmentation masks for blood and for person, and (iii) an initial classification of severe hemorrhage detection; all of this information could then be passed into ViP-LLAVA for a detailed reasoning analysis and confidence assignment. Extending the idea, a human in the loop could regularly add annotations for scene understanding (natural language annotations, filling in details that the VLM missed), as well as labeling True/False binary ground truth of severe hemorrhage. If such a pipeline could be set up as a continuously differentiable sequence of models and operations, this would lead to a learning signal that would flow back to both ViP-LLAVA for reasoning about the segmentation masks and semantic information and all the way to the original VLM (possibly vanilla LLAVA) - which generates scene descriptions. With the help of human domain experts (e.g. the combat medics on our team), such a learning system could eventually produce a highly robust model that captures the reasoning abilities of a seasoned combat medic.

We are excited to begin exploring such ideas as we prepare for the second year of DARPA's Triage Challenge.

8 Conclusion

The insights from this study and the lessons learned from the DARPA competition highlight significant opportunities to refine and extend hemorrhage detection systems. By addressing current limitations, such as the false positives caused by red artifacts and the challenges of limited data, and by exploring innovative directions like semantic grounding and multi-task learning, we aim to develop robust, explainable, and dependable models.

Given that we started with a highly limited small dataset and a dearth of approaches to this problem, it is reassuring that our broad search led to a relatively reliable pipeline that allowed us to perform reasonably at the DARPA Challenge year-1 event, scoring second out of all the academic teams. We are committed to leveraging this foundation to contribute to cutting-edge research at the intersection of biomedical engineering, AI, and robotics. Ultimately, these advancements are not aimed at merely improving performance in simulated environments; our progress here is actively paving the way for real-world deployments that can make life-saving impacts.

9 Funding

This work was funded by the DARPA Triage Challenge's funding for Team Chiron - under grant #HR00112420329.

10 Ethics

This work was conducted in adherence to and in accordance with the following IRB protocols.

11 Credit

I have a lot of people to thank for the opportunity. Firstly, my thanks to Prof. John Galeotti who trusted me with this project and took me under his wing at a point in 2023 when my only proof of AI proficiency was a MOOC certificate and a singular blogpost about stable diffusion. The success of this work would not have been possible without his constant support and eye to detail. In that vein, my heartfelt thanks to Nishanth Arun, a good friend and mentor - neither my skills nor my worldview would be the same without having known him. I thank Prof. Sebastian Scherer (aka basti), who constantly inspired us with his energy and ever smiling, cheeky optimism. I also thank Dr. Kimberly Elenberg, who I learned innumerable lessons from. I also extend my gratitude to Dr. Lenny Weiss

and our friends at UPMC and Pittsburgh EMS, without whose help and generosity our Friday field tests and simulated runs would not have been possible. I'd also like to thank the unlikely friends I made at the Airlab - particularly Omar Alama and Adi Rauniyar. Their company kept me sane through the arduous hours at the lab; and lively discussions with them kept my work in perspective. I also wish to thank the core members of Team Chiron - Mayank, Kabir, Aniket, Parv - we underwent the rites of passage inherent to the Triage Challenge, and made sure to band together for when the team needed it the most. My heartfelt thanks also to Yaoyu Hu, and Varun K - as supervisors in the project, without their willingness to go out of their way to help us learn, think, and become leaders in our own right, the team would not be a fraction of what we grew into.

I extend my heartfelt gratitude to my parents - they tolerated my relentless questioning and expensive reading habits for a couple of decades; and as it turns out, I get to indulge the exact same obsessions here and get credit for it. Here's to my friends, particularly Vishnu and Raja, who constantly reminded me of my ambitions and roots alike. In that vein, I also thank the good friends in my Pitt life - Adarsh R, Sree Dev, Amulya, Abhimanyu, Eli, Sneh, and Vratin - without whom getting through the trials and tribulations of this project would've been a lot harder.

References

[1] Risheng Wang, Tao Lei, Ruixia Cui, Bingtao Zhang, Hongying Meng, and Asoke K Nandi. Medical image segmentation using deep learning: A survey. *IET image processing*, 16(5): 1243–1267, 2022.

[2] Karl D Fritscher, Agnes Grünerbl, and Rainer Schubert. 3d image segmentation using combined shape-intensity prior models. *International Journal of Computer Assisted Radiology and Surgery*, 1:341–350, 2007.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[4] Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit Merhof. Medical image segmentation review: The success of u-net. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10076–10095, 2024. doi: 10.1109/TPAMI.2024.3435571.

[5] M Krithika Alias AnbuDevi and K Suganthi. Review of semantic segmentation of medical images using modified architectures of unet. *Diagnostics*, 12(12):3064, 2022.

[6] Jose Dolz, Christian Desrosiers, and Ismail Ben Ayed. Ivd-net: Intervertebral disc localization and segmentation in mri with a multi-modal unet. In *International workshop and challenge on computational methods and clinical applications for spine imaging*, pages 130–143. Springer, 2018.

[7] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE transactions on medical imaging*, 37(12):2663–2674, 2018.

[8] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022.

[9] Xiangyi Yan, Hao Tang, Shanlin Sun, Haoyu Ma, Deying Kong, and Xiaohui Xie. After-unet: Axial fusion transformer unet for medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3971–3981, 2022.

[10] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[11] Yanzhou Su, Jian Cheng, Murong Yi, and Haijun Liu. Fapn: Feature augmented pyramid network for polyp segmentation. *Biomedical Signal Processing and Control*, 78:103903, 2022.

[12] Zhendong Wang, Jiehua Zhu, Shujun Fu, Shuwei Mao, and Yangbo Ye. Rfpnet: Reorganizing feature pyramid networks for medical image segmentation. *Computers in biology and medicine*, 163:107108, 2023.

[13] Yonglin Yu, Haifeng Li, Hanrong Shi, Lin Li, and Jun Xiao. Question-guided feature pyramid network for medical visual question answering. *Expert Systems with Applications*, 214:119148, 2023.

[14] Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Sid Kiblawi, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Jacob Abel, et al. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nature Methods*, pages 1–11, 2024.

[15] Peilun Shi, Jianing Qiu, Sai Mu Dalike Abaxi, Hao Wei, Frank P-W Lo, and Wu Yuan. Generalist vision foundation models for medical imaging: A case study of segment anything model on zero-shot medical segmentation. *Diagnostics*, 13(11):1947, 2023.

[16] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

[17] Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, pages 1–13, 2024.

[18] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

[19] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[20] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[21] Murtadha D Hssayeni, Muayad S Croock, Aymen D Salman, Hassan Falah Al-Khafaji, Zakaria A Yahya, and Behnaz Ghoraani. Intracranial hemorrhage segmentation using a deep convolutional model. *Data*, 5(1):14, 2020.

[22] Harvendra Singh Bhadauria, Annapurna Singh, and ML Dewal. An integrated method for hemorrhage segmentation from brain ct imaging. *Computers & Electrical Engineering*, 39(5): 1527–1536, 2013.

[23] Papangkorn Inkeaw, Salita Angkurawaranon, Piyapong Khumrin, Nakarin Inmutto, Patrinee Traisathit, Jeerayut Chaijaruwanich, Chaisiri Angkurawaranon, and Imjai Chitapanarux. Automatic hemorrhage segmentation on head ct scan for traumatic brain injury using 3d deep learning model. *Computers in Biology and Medicine*, 146:105530, 2022.

[24] Mobarakol Islam, Parita Sanghani, Angela An Qi See, Michael Lucas James, Nicolas Kon Kam King, and Hongliang Ren. Ichnet: intracerebral hemorrhage (ich) segmentation using deep learning. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I 4*, pages 456–463. Springer, 2019.

[25] Zhaodong Sun and Xiaobai Li. Contrast-phys: Unsupervised video-based remote physiological measurement via spatiotemporal contrast. In *European Conference on Computer Vision*, pages 492–510. Springer, 2022.

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

[27] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022.

[28] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[29] Nikhil Keetha, Avneesh Mishra, Jay Karhade, Krishna Murthy Jatavallabhula, Sebastian Scherer, Madhava Krishna, and Sourav Garg. Anyloc: Towards universal visual place recognition. *IEEE Robotics and Automation Letters*, 2023.

[30] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6593–6602, June 2024.

[31] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature medicine*, pages 1–12, 2024.

[32] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024.

[33] Katherine L Alfred, Andrew C Connolly, Joshua S Cetron, and David JM Kraemer. Mental models use common neural spatial structure for spatial and abstract content. *Communications biology*, 3(1):17, 2020.

[34] K Anita Davamani, CR Rene Robin, D Doreen Robin, and L Jani Anbarasi. Adaptive blood cell segmentation and hybrid learning-based blood cell classification: A meta-heuristic-based model. *Biomedical Signal Processing and Control*, 75:103570, 2022.

[35] Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. Making large multimodal models understand arbitrary visual prompts. *arXiv e-prints*, pages arXiv–2312, 2023.

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