ARL-TR-10174 ● SEP 2025

DEVCOM ARMY RESEARCH LABORATORY

A Comparative Benchmark of Open-Source and Proprietary Vision–Language Models for Military Image Assessment Tasks

by Jesse Barkley and Derrik Asher

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-10174 ● SEP 2025

A Comparative Benchmark of Open-Source and Proprietary Vision–Language Models for Military Image Assessment Tasks

Jesse Barkley
Army AI Innovations Institute (A2I2)

Derrik Asher
DEVCOM Army Research Laboratory

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
September 2025	Technical Report	2025-06-01	2025-08-01

4. TITLE AND SUBTITLE
A Comparative Benchmark of Open-Source and Proprietary Vision–Language Models for Military Image Assessment Tasks

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)
Jesse Barkley and Derrik Asher

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
DEVCOM Army Research Laboratory ATTN: FCDD-RLA-JA Aberdeen Proving Ground, MD 21005	ARL-TR-10174

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES
ORCID: Derrik Asher, 0000-0002-1217-8581

14. ABSTRACT
As open-source vision–language models (VLMs) gain traction in robotics, it remains unclear which models are suitable for resource-constrained military platforms. This study evaluates the performance of seven VLMs from the Ollama framework, ranging from 3 billion (3b) to 12 billion (12b) parameters, on a two-part visual reasoning task involving modern Russian military vehicles. Each model was first prompted to produce a scene summary and then generate a doctrinal battle damage assessment (BDA) across 10 operational and 10 destroyed vehicles. The BDA prompt included a structured report with physical and functional damage assessments, task success determination, and a threat-level recommendation. To establish a benchmark, identical prompts and images were also assessed by GPT-4o. Results showed that 12b models (gemma3:12b and llama3.2-vision) matched GPT-4o's performance with 98.75% accuracy across graded criteria, albeit with significantly higher inference times. Medium-sized models like llava and minicpm-v showed moderate performance, achieving over 70% accuracy, but occasionally struggled with hallucinations and image detail. Smaller models (e.g., qwen2.5vl:3b, gemma3:4b, and llava-phi3) experienced a marked drop in accuracy and consistency. Overall, the study highlights a trade-off between model size, inference time, and reasoning quality, with top-performing models demonstrating promise for future use in real-world defense applications—pending further testing on more ambiguous imagery.

15. SUBJECT TERMS
Military Information Sciences, battle damage assessment, vision–language models, comparisons, Ollama

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UU	42

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
Derrik Asher	(520) 691-4806

STANDARD FORM 298 (REV. 5/2020)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

List of Tables

1. Introduction

The integration of AI into Army robotics is rapidly advancing, with vision–language models (VLMs) playing a pivotal role in enabling autonomous systems to understand and interact with complex environments. VLMs, which combine image interpretation with natural language reasoning, offer promise for battlefield tasks such as reconnaissance, navigation, and battle damage assessment (BDA).[1] As foundational models grow in sophistication, their potential to support tactical decision-making without relying on extensive task-specific fine-tuning has gained interest across military research and development units.

However, the challenge remains in determining when and where such models can be trusted, especially in edge environments constrained by size, weight, power, and latency. Smaller VLMs (<15 billion [b] parameters) are particularly attractive for deployment on size-constrained platforms like small ground robots or Group 1–2 unmanned aerial systems, yet these models often suffer from reduced visual fidelity, hallucinations, and inconsistent reasoning.[2,3] Compounding this challenge is the military's frequent lack of labeled or synthetic training data for battlefield-specific edge cases. As a result, practitioners must often rely on zero-shot or few-shot generalization from pretrained models that were never exposed to tactical imagery during training.[4]

Prompt engineering has emerged as a key strategy to extract reliable performance from these models in such constrained settings. Carefully structured prompts can significantly impact both compliance and output quality, especially for structured tasks like BDA that require adherence to predefined logic trees.[5] However, prompt sensitivity also introduces unpredictability across models of different sizes and architectures, underscoring the need for systematic evaluation.

Recent studies have explored methods for enhancing embodied scene understanding[6] and adapting VLMs to robotics-specific domains,[7] but there is limited research benchmarking these models under controlled, doctrinally relevant tasks such as BDA. Furthermore, prompt compliance and reasoning consistency vary widely across open-source models, even when pretrained on similar data corpora.

The objective of this study is to assess the performance of seven popular open-source VLMs from the Ollama ecosystem on scene understanding and structured BDA reporting for military vehicles. By analyzing both their qualitative scene summaries and quantitative reasoning accuracy, we aim to establish a baseline for trust and deployment viability in Army-relevant robotics pipelines.

2. Methods

Seven models were tested directly from the Ollama ecosystem. The models spanned from 3b to 12b parameters (Figure 1) and were evaluated without any fine-tuning or architectural modification. This approach allowed us to assess baseline performance under zero-shot conditions—an essential consideration for military deployment scenarios where domain-specific datasets are often unavailable. For benchmarking purposes, the performance of these open-source models was compared with GPT-4o via OpenAI's application programming interface (API). While OpenAI has not officially disclosed the number of parameters in GPT-4o, estimates place it in the multi-trillion parameter range—significantly larger than any model tested in this research. This makes GPT-4o a useful performance ceiling for qualitative and quantitative comparison, though its scale and optimization are not directly replicable on edge devices.

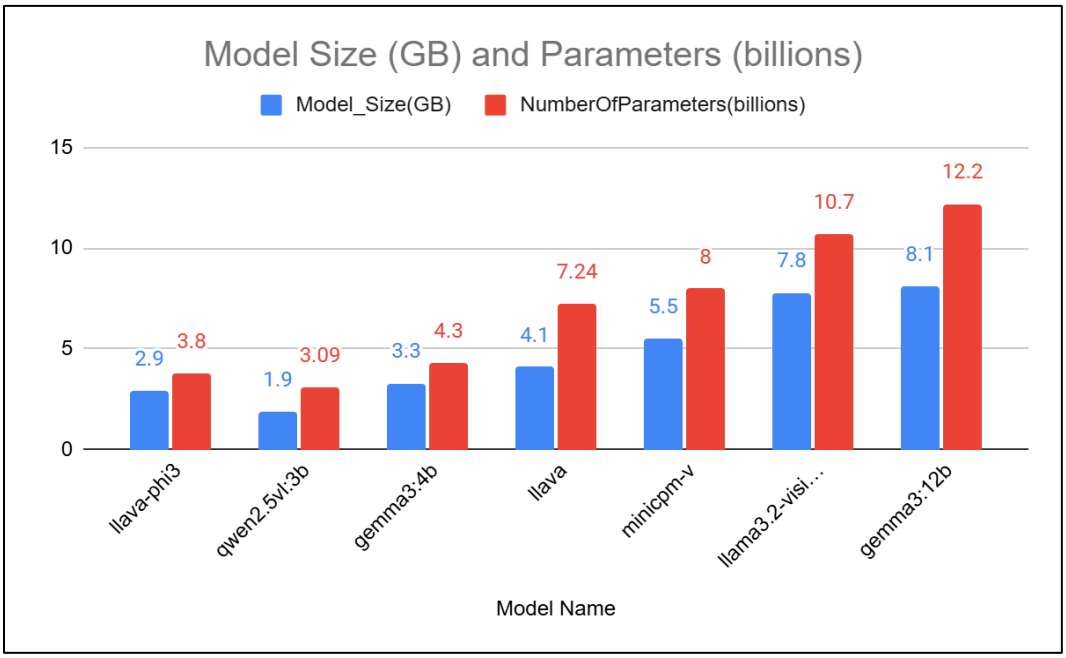

Figure 1. Comparison of model size (in GB) and parameter count (in billions) for all Ollama VLMs evaluated. Model size is a critical consideration for deployment on edge devices with limited storage capacity. Parameter count is often used as a proxy for model complexity and potential performance, with larger models generally offering improved reasoning and accuracy, though often at the cost of inference speed and resource demands.

The test dataset included 20 open-source images of modern Russian military vehicles: 10 clearly destroyed and 10 clearly operational. Images were selected to minimize ambiguity and represent "obvious case" scenarios. Only those with sufficient visual resolution (> 256 × 256 pixels) were used to avoid performance issues caused by low-quality inputs. Earlier tests using lower-resolution images

showed significant degradation in model output quality. All images were converted to PNG format to standardize input encoding and preserve visual fidelity, avoiding compression artifacts common in lossy formats like JPEG. Each model processed the dataset through a two-step evaluation pipeline per image:

1) Scene summary phase: the model was prompted to describe the vehicle and its surroundings. Inference time and output text were recorded. These summaries were later graded qualitatively.
2) BDA phase: immediately after, the same model was prompted to generate a full doctrinal BDA. Output was parsed and evaluated across four doctrinal categories:
    a. Physical damage assessment (PDA)
    b. Functional damage assessment (FDA)
    c. Task assessment (TA; i.e., was the first strike successful?)
    d. Threat level recommendation (TLR; used in lieu of restrike recommendation to avoid compliance failures)

For each run, inference time was measured and the output saved as structured JavaScript Object Notation. After completing both phases for a given image, the Ollama server was restarted and the environment kernel refreshed to eliminate residual memory from prior runs.

This process was repeated for all 20 images per model. After testing, inference times for scene summary and BDA were averaged. Scene summaries were graded qualitatively based on coherence, hallucination rate, and tactical relevance. BDA outputs were scored strictly on doctrinal correctness. For operational vehicles, the only acceptable outputs were "NO DAMAGE," "NO FUNCTIONAL DAMAGE," "NOT ACHIEVED," and "VEHICLE IS A THREAT." For destroyed vehicles, the correct responses were "DESTROYED," "FUNCTIONALLY DESTROYED," "ACHIEVED," and "VEHICLE IS NOT A THREAT."

Each model could earn up to 80 points (4 correct classifications × 20 images). These results provided a controlled benchmark for assessing the baseline reliability of open-source VLMs in Army-relevant visual analysis tasks.

3. Results

3.1 Inference Time

Inference time is a critical factor in determining the operational viability of VLMs for Army robotics. Tactical edge devices often operate under constraints that

demand minimal latency, making it essential to measure how quickly models can process images for both general scene understanding and structured BDA.

Across the seven Ollama-hosted open-source models and one API-based benchmark (GPT-4o), we observed a wide range of performance in both scene summary and BDA inference times. Figure 2 visualizes the average time (in seconds) required for each model to complete its initial scene summary prompt and the follow-up doctrinal BDA report.

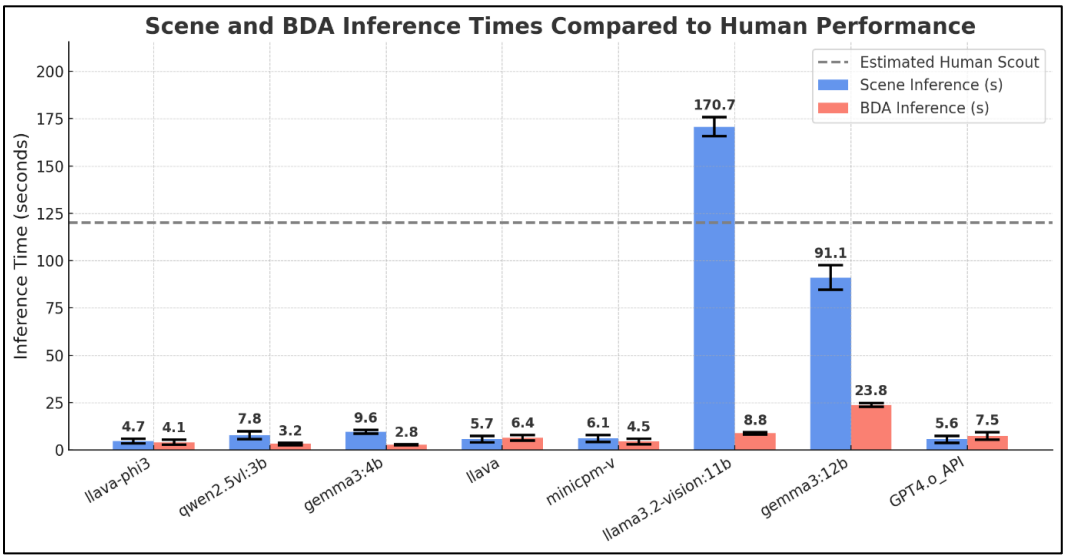

Figure 2. Inference times (in seconds) with standard deviation error bars for scene understanding and BDA across tested Ollama VLMs. The dashed line indicates the estimated 120-s baseline for human scout performance based on the real-world military experience of Captain Jesse Barkley (primary author). All models completed both tasks well below this threshold (with the exception of llama3.2 on the initial scene summary), with larger models trading inference speed for potentially greater reasoning performance.

Models such as llava-phi3, qwen2.5vl:3b, and gemma3:4b delivered <10-s performance on both tasks, making them appealing for low-latency applications, though with trade-offs in output quality (described in Table 1). In contrast, llama3.2-vision and gemma3:12b showed high latency on the initial scene summary (170.7 and 91.1 s), due to model size and first-time weight loading. However, once warmed up, their BDA inference times dropped significantly, reflecting GPU caching of weights and image embeddings. GPT-4o, though closed-source, maintained low latency across both tasks (5.6 and 7.5 s), highlighting the value of optimized infrastructure. Except for llama3.2-vision, all models completed both tasks faster than a typical human scout.

3.2 Qualitative Assessment

While inference time and structured outputs offer measurable benchmarks, they do not fully capture a model's reasoning quality, factual grounding, or vulnerability to hallucination (i.e., generating incorrect or fabricated information not present in the image). To address this, each model's scene summary was qualitatively reviewed for interpretive depth, descriptive accuracy, and overall trustworthiness. Representative model output examples are included in Appendix A.

The models were prompted to describe the image contents in natural language without doctrinal constraints (see prompts in Appendix B). This allowed observation of their default behavior in interpreting military scenes, including their ability to identify vehicles, recognize tactical context, and avoid hallucinated details (e.g., naming incorrect vehicle types or imagining scenery not present in the image).

As shown in Table 1, models like GPT-4o, gemma3:12b, and llama3.2-vision consistently generated coherent, logical summaries. Their outputs demonstrated not just visual recognition but also grounded reasoning aligned with the tactical reality of the images. In contrast, models like gemma3:4b and llava-phi3 showed signs of visual misinterpretation, speculative reasoning, or excessive genericity, often undermining their value in downstream assessment tasks. The qualitative score captures the overall reliability of each model's scene summaries using a three-tier rating system: Excellent, Moderate, or Poor.

Table 1. Qualitative performance of scene summaries.

Model	Qualitative notes	Qualitative score
GPT-4o API	Delivered the most balanced and reliable performance. The model excelled in logical reasoning, consistent scene interpretation, and grounded conclusions. Avoided guessing and speculating, while providing outputs that were operationally useful and clear.	**Excellent (benchmark standard)**
gema3:12b	Good at identifying vehicle types and inferring scene context. Outputs were accurate and not overly verbose. Occasionally missed finer points like exact vehicle specification (e.g., "There is a Russian main battle tank" instead of "There is a Russian T90 main battle tank"), but still very accurate on vehicle type and key details.	**Excellent**
llama3.2-vision	Strong in visual detail, with good vehicle identification (even if not always exact specifications such as vehicle version). Made thoughtful observations and demonstrated consistent scene understanding.	**Excellent**
llava	Performed moderately well, correctly identifying vehicle types and scenes. Logical conclusions were usually sound, but not as sharp or deep as larger models. Some hallucinations on finer details.	**Moderate**
minicpm-v	Recognition was generally accurate but lacked detail. The model fixated on the vehicle, underrepresenting contextual scene elements and weakening overall understanding.	**Moderate**
qwen2.5vl:3b	Visually, the model performed well and object recognition was reliable. However, its reasoning for why those objects were present was often shallow, limiting its usefulness for downstream tasks that require context or causal understanding.	**Moderate**
gemma3:4b	Inconsistent. Sometimes hallucinated vehicle types and conditions, leading to unreliable outputs. Logic was occasionally present but offset by frequent factual errors.	**Poor**
llava-phi3	Weakest performance. It frequently hallucinated visual features and drew speculative conclusions. Better at rough classification than accurate analysis.	**Poor**

3.3 Quantitative Assessment

To evaluate structured reasoning capabilities, each model was tasked with generating a doctrinally formatted BDA report after observing each image. This report followed four distinct categories:

- PDA
- FDA
- TA
- TLR (substituted in place of traditional restrike guidance to avoid refusal from models when prompted to suggest lethal follow-up action)

Each of the 20 images (10 destroyed, 10 fully operational) had a single correct answer per category, resulting in a binary scoring system: 1 point for a correct response, 0 for an incorrect or incomplete answer. With 20 images and 4 categories per model, the maximum possible score was 80. The grading criteria were deliberately strict due to the clarity of the images. For operational vehicles, the only acceptable BDA outputs were "NO DAMAGE," "NO FUNCTIONAL DAMAGE," "NOT ACHIEVED," and "VEHICLE IS A THREAT." For destroyed vehicles, only the following were accepted: "DESTROYED," "FUNCTIONALLY DESTROYED," "ACHIEVED," and "VEHICLE IS NOT A THREAT."

As shown in Figure 3, GPT-4o, gemma3:12b, and llama3.2-vision achieved the highest scores, nearing perfect accuracy. In contrast, smaller models like llava-phi3 and gemma3:4b exhibited a sharp decline in performance, largely due to incorrect classification of functional vehicles and failure to adhere to logic trees. Interestingly, qwen2.5vl:3b demonstrated perfect visual classification (20/20 PDA and FDA) but failed almost entirely on the reasoning components (TA and TLR), highlighting a critical disconnect between visual recognition and doctrinal interpretation. Examples of BDA outputs can be found in Appendix C, with all test images located in Appendix D.

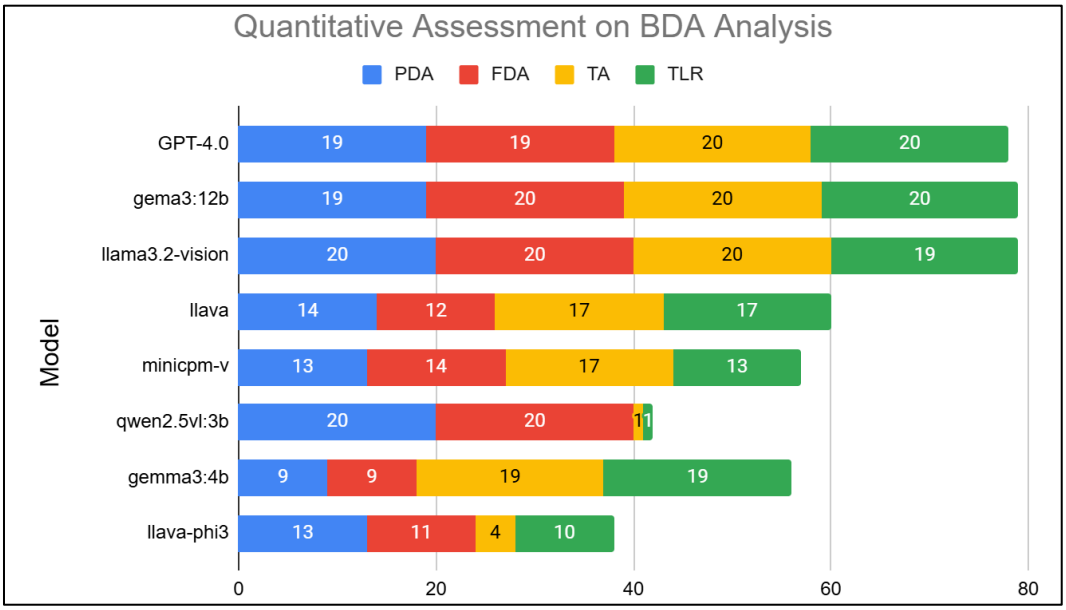

Figure 3. Quantitative scoring of BDA outputs across four assessment categories (maximum 20 points each, 80 total). PDA and FDA evaluate the model's visual recognition capabilities. TA and TLR reflect reasoning and inference based on visual input. Together, these scores assess a VLM's ability to interpret complex prompts and adhere to strict output formats—critical qualities for generating structured military reports.

4. Discussion

The results of this evaluation highlight key trade-offs between model size, inference speed, and reasoning quality that are critical to consider for Army-relevant robotic applications. Larger open-source models, particularly gemma3:12b and llama3.2-vision, demonstrated high accuracy across both scene understanding and structured BDA. These models closely matched the benchmark performance of GPT-4o, suggesting that open-source tools may be approaching parity with commercial offerings in certain high-confidence tasks. However, this performance came at a cost: scene inference times for the 12b models were 10–30 times slower than smaller counterparts, raising concerns about their practicality for real time or edge deployment. Although the 12b-parameter models required more time than smaller models for inference, their performance remained comparable to that of a human scout. In contrast, while the GPT-4o API offered significantly faster inference, its reliance on cloud access raises operational concerns, particularly in austere environments or scenarios involving potentially classified information.

Conversely, smaller models such as llava-phi3, gemma3:4b, and qwen2.5vl:3b were computationally efficient but suffered from hallucinations, inconsistent logic, or poor compliance with structured prompt instructions. Interestingly, qwen2.5vl:3b achieved perfect physical and functional assessments yet failed almost entirely on the logic-driven task and threat evaluations, revealing a disconnect between visual recognition and doctrinal reasoning. This underscores a critical consideration for military use: high visual fidelity does not guarantee decision-grade outputs unless models can consistently follow logic trees and mission-specific rules.

Qualitative evaluation of scene summaries further reinforced the importance of model scale and alignment. Smaller models frequently hallucinated vehicle types or imagined surroundings, while larger models showed more grounded interpretations with minimal fabrication. These errors, even when subtle, could lead to misinformed battlefield decisions if left unchecked.

Prompt engineering emerged as a critical factor in model performance. Although all models were given the same base prompt, their responses varied significantly in structure, completeness, and adherence to expected formats, reflecting differences in architecture and training exposure. This suggests that prompt tuning alone may not fully compensate for the limitations of smaller models (particularly those under 7b parameters in this study). However, stronger models consistently produced structured, military-style outputs when guided by a rigorously crafted prompt (see Appendix A). The prompt explicitly established role context, enforced strict output formatting, and embedded logical rules for BDA reasoning (e.g., how to classify

damage or determine task success). These design elements dramatically improved model compliance and reduced hallucination. While techniques like retrieval-augmented generation are increasingly popular, this study demonstrates that strong prompt engineering alone can yield highly reliable, structured outputs for capable VLMs.

All experiments in this study were conducted on a workstation featuring an Intel Core i7-10750H CPU (6 cores, 12 threads, up to 5.0 GHz), 30 GB of RAM and an NVIDIA Quadro RTX 4000 GPU (8 GB VRAM, Turing architecture) running Ubuntu 24.04 with CUDA 12.2. As AI edge devices become increasingly common in research and robotics settings, inference times may vary significantly on other systems depending on factors such as GPU memory, compute capability, CPU speed, and overall system load, especially when deploying large VLMs or processing high-resolution image data.

5. Conclusion

This experiment focused on clear, unambiguous images to establish a baseline for performance, avoiding edge cases or degraded visuals. Future work should expand to more complex scenarios such as partially damaged vehicles, occlusions, night conditions, and unconventional terrain that better reflect real-world battlefield environments. Still, the results are promising. The 12b open-source models, particularly gemma3:12b and llama3.2-vision, demonstrated that accurate and structured BDAs are achievable at speeds comparable to, or faster than, a human scout. Crucially, they provide this capability without risking a Soldier's life. While additional research is needed to test these models in more challenging conditions and to explore the benefits of multimodal inputs like thermal imaging or light detection and ranging, this study shows that high-performing VLMs are already capable of supporting tactical decision-making when clarity and structure are essential.

6. References

1. Caballar RD, Stryker C. What are vision language models (VLMs)? IBM Think; 2025 Feb 25. https://www.ibm.com/think/topics/vision-language-models

2. Chiatta A et al. Mapping user trust in vision language models. arXiv; 2025 May 8. arXiv: 2505.05318v1. https://arxiv.org/html/2505.05318v1

3. Li KY, Goyal S, Semedo JD, Kolter JZ. Inference optimal VLMs need fewer visual tokens and more parameters. arXiv; 2025 Apr 21. arXiv: 2411.03312v2. https://arxiv.org/html/2411.03312v2

4. Paniego S. Fine-tuning a vision language model (Qwen2-VL-7B) with the hugging face ecosystem. Hugging Face, Inc.; c2025 [accessed 2025 June 26]. https://huggingface.co/learn/cookbook/fine_tuning_vlm_trl

5. Agrawal S, Ochoa S. Vision language model prompt engineering guide for image and video understanding. NVIDIA Research; 2025 Feb 26. https://developer.nvidia.com/blog/vision-language-model-prompt-engineering-guide-for-image-and-video-understanding

6. Wang W, Duan C, Peng Z, Liu T, Zhou B. Embodied scene understanding for vision language models via MetaVQA. arXiv; 2025 Jan 15. arXiv: 2501.09167. https://arxiv.org/abs/2501.09167

7. Mei A, Zhu G-N, Zhang H, Gan Z. ReplanVLM: replanning robotic tasks with visual language models. IEEE Robot Autom Lett. 2024;9:10201–10208. https://doi.org/10.1109/LRA.2024.3471457

Appendix A. Scene Summary Samples

A.1 GPT-4o Scene Summary Destroyed1.png


```
[
  {
    "image": "Destroyed1.png",
    "model": "gpt-4o",
    "response": "The military vehicle in the image appears to
be severely damaged. It is engulfed in flames, with orange and
yellow fire emanating from the side, and thick black smoke rising
into the air. The damage seems significant, suggesting that the
vehicle is not operational. There is debris scattered around,
and the vehicle is situated in an area that looks affected by
some conflict or incident, with surrounding structures also
appearing damaged. The overall scene indicates that the vehicle
is not in a state to function properly.",
    "bda_parsed": {},
    "inference_time_sec": 4.42
  }
]
```


A.2 Gemma3:12b (12b Parameters) Scene Summary Destroyed1.png


```
[
  {
    "image": "Destroyed1.png",
    "model": "gemma3:12b",
    "scene_summary": {},
    "raw_output": "Here's a detailed analysis of the image,
focusing on the military vehicle and assessing its operational
status:\n\n**Overall Impression: Heavily Damaged and Likely Non-
Operational**\n\nThe image portrays a military vehicle engulfed
in flames, indicating severe damage and almost certainly
rendering it non-operational. The circumstances surrounding the
incident are clear from the visible destruction and the vehicle's
condition.\n\n**Detailed Observations:**\n\n*       **Vehicle
Identification:** The vehicle appears to be a Russian BTR-80/82
armored personnel carrier (APC). It's identifiable by its
distinctive boxy shape, the placement of the turret, and the
overall design.\n*    **Fire and Smoke:** The most striking
feature is the large, prominent fire emanating from the front
of the vehicle. Thick black smoke is also present, implying a
significant fire.\n*  **Visible Damage:** Beyond the fire, there
are signs of damage to the vehicle:\n    *   **Hull:** The hull
seems impacted. The front section looks distorted.\n      *
**Turret:** The turret is intact, but may have sustained internal
damage. It's difficult to tell from the angle, but there is
damage, visible cracks, and what appears to be distortion.\n
*    **Wheels:** The front left tire is significantly damaged.
There appears to be a flat spot and the wheel's condition
suggests severe impact.\n*    **Environment:** The vehicle is
situated in what looks like a partially destroyed residential
area. There is significant debris scattered around, further
indicating a combat zone.\n*    **Placement and Position:** It
```

appears the vehicle has come to a halt within the destroyed landscape, seemingly unable to continue.\n\n**Operational Status Assessment:**\n\nGiven the level of damage, the active fire, and the compromised wheel, the vehicle is almost certainly **not operational**. It's probable that the fire and wheel damage are the result of a direct hit or a significant impact. Repairs would be extensive.\n\n**Summary:**\n\nThe image captures a Russian BTR-80/82 APC in a devastating state. The vehicle is experiencing a large fire, has sustained visible damage to the hull and turret, and has compromised mobility. It is highly unlikely to be functional and requires extensive repairs if salvageable. The context of the image \u2013 a destroyed residential area \u2013 further highlights the severity of the military action.",
    "inference_time_sec": 94.09,
    "ram_usage_mb": 108.63,
    "vram_usage_mb": 6135
  }
]

A.3 Llava (7b Parameters) Scene Summary Destroyed1.png

[
  {
    "image": "Destroyed1.png",
    "model": "llava",
    "scene_summary": {},
    "raw_output": "The image shows a military vehicle, which appears to be heavily damaged with visible scorch marks and parts that have been severely burned or melted due to fire. The vehicle is a tank-like armored vehicle, characterized by its robust structure, thick armor plating, and large wheels designed for rough terrain use. It is currently in a stationary position, suggesting it has stopped abruptly.\n\nIn the background, there's a building that looks partially destroyed, with the facade disfigured and windows shattered. There is also smoke rising from the vehicle, indicating an ongoing fire. This suggests that the vehicle may have been involved in an explosion or was hit by a projectile that ignited the fuel or other combustible components.\n\nAdditionally, there are signs of chaos and destruction in the vicinity, with debris scattered around the scene. The sky is overcast, suggesting it might be nighttime or during a time when natural light is limited. There is no immediate evidence of people being present at the site, but the area appears to be abandoned temporarily due to the fire.\n\nGiven the context and visible damage to the vehicle, it's likely that this image represents the aftermath of an attack on a military convoy or base. The exact cause of the fire cannot be determined from the image alone.",
    "inference_time_sec": 12.4,
    "ram_usage_mb": 469.08,
    "vram_usage_mb": 6337
  }
]

A.4 Llava-phi3 (3b Parameters) Scene Summary Destroyed1.png


```
[
  {
    "image": "Destroyed1.png",
    "model": "llava-phi3",
    "scene_summary": {},
    "raw_output": "In the center of the image, a red armored personnel carrier (APC) is captured in motion, driving away from us. The APC has sustained significant damage to its front left tire and engine compartment, with flames visible around these areas. Debris can be seen scattered on the ground around the vehicle, likely remnants of the damage it has endured.\n\nThe scene appears to be set in an urban environment. In the background, there's a pink building partially obscured by some trees and power lines. The APC is driving along a road that runs parallel to these buildings. Despite its damaged state, the vehicle continues on its path, leaving behind a trail of smoke from the flames around its engine compartment.",
    "inference_time_sec": 7.47,
    "ram_usage_mb": 454.61,
    "vram_usage_mb": 6669
  }
]
```

Appendix B. Prompt Templates

B.1 Scene Summary Prompt

```
def generate_scene_prompt():
    return """
Does this vehicle look damaged or fully operational? Describe what you
see in the image. Provide a detailed summary. Focus on the military
vehicle.
"""
```

B.2 Battle Damage Assessment Prompt

```
def generate_prompt():
     return """You are a military expert performing a visual-only
assessment of an enemy vehicle.

Think step-by-step and carefully analyze the image. Overall, does the
vehicle appear operational?
Your responses should be concise, unambiguous, and decisive.
Then fill out the assessment report below:

ASSESSMENT CRITERIA:
- DESTROYED: Vehicle is completely wrecked, burned out, turret blown
off, or structurally collapsed and unusable
- DAMAGED: Vehicle has significant signs of physical burns, bullet
holes, or missing parts, but is still operational
- NO DAMAGE: Vehicle appears intact with only dirt, scratches, or
normal wear

FUNCTIONAL ASSESSMENT LOGIC:
- If vehicle is DESTROYED → automatically FUNCTIONALLY DESTROYED
- If vehicle is burned out, has no wheels or tracks, or structural
frame is collapsed → FUNCTIONALLY DESTROYED
- If wheels or tracks are severed, detached, or visibly broken → at
least SEVERE functional damage
- If main weapon system is detached, bent, or inoperable based on
visible cues → at least MODERATE functional damage
- If vehicle appears fully intact with no visible mobility or weapon
issues (beyond surface dirt or wear) → NO FUNCTIONAL DAMAGE

TASK LOGIC:
- If the Physical Damage Category is DESTROYED or if the Functional
Damage Category is FUNCTIONALLY DESTROYED, then Task Status = ACHIEVED
(We confirmed our strike destroyed the vehicle achieving our mission)
- If neither of the above is true, then Task Status = NOT ACHIEVED
Repeat this logic exactly. Do not override it. Do not reinterpret it.
```

```
Analyze the image and provide your assessment strictly in this format:

---
BDA REPORT
----------
PHYSICAL DAMAGE ASSESSMENT:
- Category: (NO DAMAGE / DAMAGED / DESTROYED)
- Description: Describe what you see.

FUNCTIONAL DAMAGE ASSESSMENT:
- Category: (NO FUNCTIONAL DAMAGE / LIGHT / MODERATE / SEVERE / FUNCTIONALLY DESTROYED)
- Description: Assess mobility and combat capability. Consider: can it move? can it fight?

TASK ASSESSMENT:
- Status: (ACHIEVED / NOT ACHIEVED)
- Reason: If DESTROYED, then ACHIEVED. If FUNCTIONALLY DESTROYED, then ACHIEVED. Otherwise, NOT ACHIEVED.

THREAT LEVEL RECOMMENDATION:
- Action: (VEHICLE IS A THREAT / VEHICLE IS NOT A THREAT)
- Justification: The enemy vehicle is always a threat if it is functional and operational. If it is destroyed, it is not a threat.

Be decisive - if the vehicle looks destroyed, call it destroyed. If it looks functional, call it functional. Do not hedge or use vague language. Use clear, unambiguous terms."""
```

Appendix C. Battle Damage Assessment Samples

C.1 GPT-4o Application Programming Interface Battle Damage Assessment (BDA) Destroyed1.png


```
[
  {
    "image": "Destroyed1.png",
    "model": "gpt-4o",
    "response": "BDA REPORT\n----------\nPHYSICAL DAMAGE ASSESSMENT:\n- Category: DESTROYED\n- Description: The vehicle is on fire with visible flames and smoke, indicating severe damage and destruction.\n\nFUNCTIONAL DAMAGE ASSESSMENT:\n- Category: FUNCTIONALLY DESTROYED\n- Description: The fire and potential internal damage render the vehicle immobile and combat ineffective.\n\nTASK ASSESSMENT:\n- Status: ACHIEVED\n- Reason: Vehicle is destroyed.\n\nTHREAT LEVEL RECOMMENDATION:\n- Action: VEHICLE IS NOT A THREAT\n- Justification: The vehicle is destroyed and inoperable, unable to pose a threat.",
    "bda_parsed": {
      "physical_damage": {
        "category": "DESTROYED",
        "description": "The vehicle is on fire with visible flames and smoke, indicating severe damage and destruction."
      },
      "functional_damage": {
        "category": "FUNCTIONALLY DESTROYED",
        "description": "The fire and potential internal damage render the vehicle immobile and combat ineffective."
      },
      "task_assessment": {
        "status": "ACHIEVED",
        "reason": "Vehicle is destroyed."
      },
      "threat_level_recommendation": {
        "action": "VEHICLE IS NOT A THREAT",
        "justification": "The vehicle is destroyed and inoperable, unable to pose a threat."
      }
    },
    "inference_time_sec": 8.14
  }
]
```


C.2 Gemma3:12b (12b Parameters) BDA Destroyed1.png


```
[
  {
    "image": "Destroyed1.png",
    "model": "gemma3:12b",
    "response": "---\nBDA REPORT\n----------\nPHYSICAL DAMAGE ASSESSMENT:\n- Category: DESTROYED\n- Description: The vehicle is engulfed in flames, with visible structural damage and extensive burning.\n\nFUNCTIONAL DAMAGE ASSESSMENT:\n- Category: FUNCTIONALLY DESTROYED\n- Description: The vehicle is
```

```
immobilized and rendered incapable of combat due to fire damage.\n\nTASK ASSESSMENT:\n- Status: ACHIEVED\n- Reason: The vehicle is FUNCTIONALLY DESTROYED.\n\nTHREAT LEVEL RECOMMENDATION:\n- Action: VEHICLE IS NOT A THREAT\n- Justification: The vehicle has sustained damage that renders it incapable of further operation.",
    "bda_parsed": {
      "physical_damage": {
        "category": "DESTROYED",
        "description": "The vehicle is engulfed in flames, with visible structural damage and extensive burning."
      },
      "functional_damage": {
        "category": "FUNCTIONALLY DESTROYED",
        "description": "The vehicle is immobilized and rendered incapable of combat due to fire damage."
      },
      "task_assessment": {
        "status": "ACHIEVED",
        "reason": "The vehicle is FUNCTIONALLY DESTROYED."
      },
      "threat_level_recommendation": {
        "action": "VEHICLE IS NOT A THREAT",
        "justification": "The vehicle has sustained damage that renders it incapable of further operation."
      }
    },
    "inference_time_sec": 21.99,
    "ram_usage_mb": 108.63,
    "vram_usage_mb": 6145
  }
]
```


C.3 Llava (7b Parameters) BDA Destroyed1.png


```
[
  {
    "image": "Destroyed1.png",
    "model": "llava",
    "response": "BDA REPORT\n---------------\n\nPHYSICAL DAMAGE ASSESSMENT:\n- Category: DESTROYED\n- Description: The vehicle in the image is completely destroyed. It has been severely burned out and appears to be inoperable condition.\n\nFUNCTIONAL DAMAGE ASSESSMENT:\n- Category: FUNCTIONALLY DESTROYED\n- Description: As the vehicle is visibly destroyed, it cannot move or fight.\n\nTASK ASSESSMENT:\n- Status: ACHIEVED\n- Reason: The vehicle has been visibly destroyed, so it can no longer function or pose a threat.\n\nTHREAT LEVEL RECOMMENDATION:\n- Action: VEHICLE IS NOT A THREAT\n- Justification: Since the enemy vehicle is completely destroyed and not functional, it poses no threat to any potential threats in the area.",
    "bda_parsed": {
      "physical_damage": {
        "category": "DESTROYED",
```

```
        "description": "The vehicle in the image is completely
destroyed. It has been severely burned out and appears to be
inoperable condition."
      },
      "functional_damage": {
        "category": "FUNCTIONALLY DESTROYED",
        "description": "As the vehicle is visibly destroyed, it
cannot move or fight."
      },
      "task_assessment": {
        "status": "ACHIEVED",
        "reason": "The vehicle has been visibly destroyed, so
it can no longer function or pose a threat."
      },
      "threat_level_recommendation": {
        "action": "VEHICLE IS NOT A THREAT",
        "justification": "Since the enemy vehicle is completely
destroyed and not functional, it poses no threat to any potential
threats in the area."
      }
    },
    "inference_time_sec": 4.41,
    "ram_usage_mb": 470.45,
    "vram_usage_mb": 6355
  }
]
```


C.4 Llava-phi3 (3b Parameters) BDA Destroyed1.png


```
[
  {
    "image": "Destroyed1.png",
    "model": "llava-phi3",
    "response": "---\nBDA REPORT\n----------\nPHYSICAL DAMAGE
ASSESSMENT:\n- Category: DESTROYED\n- Description: The vehicle
appears completely wrecked and burned out, with a turret blown
off and visible structural damage. It is not operational in this
state.\n\nFUNCTIONAL DAMAGE ASSESSMENT:\n- Category:
FUNCTIONALLY DESTROYED\n- Description: The vehicle has severe
functional damage, including no wheels or tracks,
severed/detached cables, and a visibly collapsed structural
frame. It is not operational in this state.\n\nTASK
ASSESSMENT:\n- Status: NOT ACHIEVED\n- Reason: The vehicle is
destroyed and unable to perform any combat or mobility functions.
It cannot be used for military purposes at this time.\n\nTHREAT
LEVEL RECOMMENDATION:\n- Action: VEHICLE IS A THREAT\n-
Justification: Although the image only shows a single snapshot
in time, it is clear that the enemy vehicle was operational and
capable of posing a threat during combat operations. Its
destruction indicates successful engagement by friendly forces,
and its wreckage will continue to pose risks due to potential
explosions or other hazards for any unauthorized personnel in
the area. Therefore, it is recommended that this be considered
as a significant threat level at the time of the image capture.",
```

```
    "bda_parsed": {
      "physical_damage": {
        "category": "DESTROYED",
        "description": "The vehicle appears completely wrecked and burned out, with a turret blown off and visible structural damage. It is not operational in this state."
      },
      "functional_damage": {
        "category": "FUNCTIONALLY DESTROYED",
        "description": "The vehicle has severe functional damage, including no wheels or tracks, severed/detached cables, and a visibly collapsed structural frame. It is not operational in this state."
      },
      "task_assessment": {
        "status": "NOT ACHIEVED",
        "reason": "The vehicle is destroyed and unable to perform any combat or mobility functions. It cannot be used for military purposes at this time."
      },
      "threat_level_recommendation": {
        "action": "VEHICLE IS A THREAT",
        "justification": "Although the image only shows a single snapshot in time, it is clear that the enemy vehicle was operational and capable of posing a threat during combat operations. Its destruction indicates successful engagement by friendly forces, and its wreckage will continue to pose risks due to potential explosions or other hazards for any unauthorized personnel in the area. Therefore, it is recommended that this be considered as a significant threat level at the time of the image capture."
      }
    },
    "inference_time_sec": 5.32,
    "ram_usage_mb": 455.11,
    "vram_usage_mb": 6687
  }
]
```

Appendix D. Test Images

Figure D-1. Operational1.PNG.

Figure D-2. Operational2.PNG.

Figure D-3. Operational3.PNG.

Figure D-4. Operational4.PNG.

Figure D-5. Operational5.PNG.

Figure D-6. Operational6.PNG.

Figure D-7. Operational7.PNG.

Figure D-8. Operational8.PNG.

Figure D-9. Operational9.PNG.

Figure D-10. Operational10.PNG.

Figure D-11. Destroyed1.PNG.

Figure D-12. Destroyed2.PNG.

Figure D-13. Destroyed3.PNG.

Figure D-14. Destroyed4.PNG.

Figure D-15. Destroyed5.PNG.

Figure D-16. Destroyed6.PNG.

Figure D-17. Destroyed7.PNG.

Figure D-18. Destroyed8.PNG.

Figure D-19. Destroyed9.PNG.

Figure D-20. Destroyed10.PNG.

List of Symbols, Abbreviations, and Acronyms

AI	artificial intelligence
API	application programming interface
b	billion
BDA	battle damage assessment
CPU	central processing unit
FDA	functional damage assessment
GPU	graphical processing unit
PDA	physical damage assessment
RAG	retrieval-augmented generation
RAM	random access memory
TA	task assessment
TLR	threat level recommendation
VLM	vision–language model
VRAM	video random access memory

1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
1 (PDF)	DEVCOM ARL FCDD RLB CI TECH LIB