Generative AI Models, Agents and Tools for Multimodal Training Scenario Generation and Mission Planning

Will Dupree, Svitlana Volkova, Hsien-Te Kao, Grant Engberson, Miles Markey, Gabe Ganberg, Alexxa Bessey, Summer Rebensky

Aptima, Inc.

Woburn, Massachusetts

{wdupree, svolkova, hkao, gengberson, mmarkey, gganberg, abessey, srebensky}@aptima.com

Thomas Dubai
Nikola Cardenas

Combat Air Force (CAF) Distributed Training Center (DTC)

Langley AFB Hampton VA

{Thomas.dubai.ctr, Nikola.cardenas.1.ctr}@us.af.mil

ABSTRACT

The Department of Defense (DoD) and the United States Air Force rely on computer-generated forces (CGFs) to reduce costs and resources in fighter operator training. Software such as the Next Generation Threat System (NGTS) enables complex tactical engagement simulations using a red vs. blue wargaming paradigm (e.g., 2-v-2 dogfights or coastal attack and defense engagements). However, the number of large-scale scenarios necessary for advanced operator training is limited. Key obstacles in preparing scenarios for wargaming in NGTS include the time and expertise required to devise and script complex scenarios, as well as the challenge of transcribing this technical knowledge to the machine via graphical user interfaces (GUIs) and software engineer support. In this work, we introduce a novel compound AI framework that leverages large language models (LLMs), vision language models (VLMs) and agentic AI workflows augmented with structured and unstructured data representations to revolutionize multimodal training scenario generation, retrieval and summarization within the NGTS ecosystem. Our ForceGen approach implements a three-stage pipeline: (1) multimodal knowledge extraction using semantic chunking algorithms and VLM-based image analysis to process 11,237 DTC files totaling 39.7GB; (2) hybrid Retrieval-Augmented Generation (RAG) architecture combining sentence-transformers with image and table encodings; and (3) structured scenario generation producing CGF-compliant XML with validated schema adherence. The system successfully extracts and represents complex tactical relationships including aircraft positioning, weapon loadouts, engagement zones, and communication infrastructure from multimodal inputs. The system's CGF-agnostic architecture enables seamless translation between NGTS, ASCOT 7, and emerging JSE platforms, directly addressing the critical operational requirement for rapid adversary TTP adaptation. This capability ensures Combat Air Force units can refresh training scenarios daily based on current intelligence, fundamentally altering the readiness generation paradigm for distributed mission operations and maintaining tactical overmatch against peer adversaries leveraging similar AI-enhanced training systems.

ABOUT THE AUTHORS

Dr. Will Dupree, Senior Research Engineer, Aptima, Inc., serves as Data Scientist Lead in the Intelligent Performance Analytics Division. He has led and contributed to multiple research efforts for DoD agencies such as the National Geospatial-Intelligence Agency, Air Force, and Army, leveraging his research skills in the domains of machine learning and AI. Dr. Dupree has been the PI on multiple SBIR-funded efforts focused on improving analytical capabilities through innovative applications of machine learning and AI. His research interests include deep learning for time series analysis, network graph modeling, and causal inference. Previously, Dr. Dupree led a team in developing advanced techniques to recognize patterns of life using satellite metadata in the space domain, with findings shared at leading conferences including the AMOS Conference. Currently, he is focused on developing tools that perform data collection and leverage graph analytics to characterize patterns found in publicly available geospatial/movement information. He utilizes his background in physics, applied mathematics, and advanced analytics to tackle complex challenges at the intersection of machine learning and defense applications. Dr. Dupree holds a PhD in physics from Washington State University and a BS in physics and applied mathematics from Montana State University.

Dr. Svitlana Volkova, Chief of AI, Science and Technology, Aptima, Inc., is a recognized leader in the field of human-centered Artificial Intelligence (AI). Her scientific leadership and outstanding research profile cover a range of topics, on natural language processing (NLP), machine learning (ML), deep learning (DL), AI test and evaluation, computational social science, and causal discovery. Dr. Volkova has served as principal investigator on 10+ DoD and DOE-funded projects (e.g., for DARPA, IARPA, NNSA) focusing on advancing various aspects of AI for national security. She leads the development of AI-powered descriptive, predictive, and prescriptive decision intelligence to model and explain complex systems and behaviors to address national security challenges in the human domain and beyond. Dr. Volkova has authored 100+ peer-reviewed

conference and journal publications. She serves as senior PC member and area chair for top-tier AI conferences and journals including AAAI, WWW, NeurPS, ACL, EMNLP, ICWSM, PNAS, and Science Advances and as a senior board member for Women in Machine Learning. Dr. Volkova is regularly invited to present her research at universities and leading tech companies such as Google Research, Facebook, and Microsoft Research. She received her PhD in computer science from The Johns Hopkins University, where she was affiliated with the Center for Language and Speech Processing and the Human Language Technology Center of Excellence.

Dr. Hsien-Te Kao, Associate Research Engineer, Aptima, Inc., is a multidisciplinary specialist in development, analysis, and evaluation in the Intelligent Performance Analytics division. He has expertise in information processing, analysis, and resilience in online discourse, focusing on digital communication and persuasive strategies. Mr. Kao has extensive experience in human-human teaming, communication traits, and team performance, with a focus on improving team dynamics and enhancing collaboration. His current leading effort is in human-AI teaming, with a focus on optimizing communication to enhance collaboration, coordination, and overall performance. Mr. Kao emphasizes the critical role of communication, interaction, and perception to achieve optimal outcomes, whether in online communication, human-human collaboration, or human-AI teaming. Dr. Kao holds PhD in computer science from University of Southern California and received his BS in mathematics from California State Polytechnic University, Pomona.

Mr. Grant Engberson, Associate Research Engineer, Aptima, Inc., specializes in artificial intelligence and brain-machine integration. Recent areas of research include computer vision, bio-signal processing, natural language processing, prompt engineering, synthetic data evaluation, timeseries forecasting, and reinforcement learning. His expertise in neurological pathophysiology provides a grounded understanding of how both humans and machines learn. Mr. Engberson received an MS in biomedical engineering from Northwestern University, and a BS in chemical and biochemical engineering from the Colorado School of Mines.

Mr. Miles Markey, Research Engineer, Aptima, Inc., works in Aptima's Intelligent Performance Analytics Division, where he uses his experience in design, development, evaluation, and implementation of human-centered machine learning systems for solving complex problems, as well as in experiment design and statistical analysis for human-centered AI applications. He also has experience in development and evaluation of AI models for computational pathology and a strong background in the healthcare space. Mr. Markey received an MS in biomedical engineering from Rutgers University and a BS in biomedical engineering from the University of Rochester.

Mr. Gabriel Ganberg, Chief Architect, Aptima, Inc. has more than 20 years of experience leading and architecting software projects in the R&D space. His focus is on applying the latest advancements in AI to solve real-world problems and on leading the development of common platforms that support Aptima's cutting-edge R&D programs. Mr. Ganberg architected and serves as technical lead for multiple Aptima platform technologies including the AI Toolbox (large language models [LLMs], agents, analytics, retrieval-augmented generation [RAG], etc.), CRAFT (long-form document and multi-dimensional analysis generation through LLMs), Discourse (generative testbed for conversation analysis and simulation), HAT (human-AI teaming), and YAADA (data-engineering infrastructure and analytic framework). Mr. Ganberg builds software in a variety of domains, including intelligence analysis, business intelligence, cyber analytics, information operations, training systems, human/AI collaboration, recommendation systems, and experimental team research. He develops tech stacks that leverage cloud architecture, distributed systems, document/graph/vector databases, machine learning, and generative AI. Mr. Ganberg received a BA in computer science and economics from Vassar College.

Dr. Alexxa Bessey, Scientist, Aptima, Inc. has a particular interest in the integration and use of technology to optimize training and performance in both individuals and teams. At Aptima, Dr. Bessey is involved in several projects including the assessment of proficiency-based training, the evaluation of simulator fidelity and sim-based training, the use of technology-based unobtrusive measures in teams, the examination of social networks and resilience, and leadership interventions. Dr. Bessey has more than 9 years of experience conducting military research to include her role as an operational research psychologist at the Walter Reed Army Institute of Research (WRAIR), in which she led two research projects with Special Operations Forces (SOF) operators. Dr. Bessey received an MPS in clinical psychological sciences from the University of Maryland College Park an MS in industrial-organizational psychology from Clemson University, and a PhD in industrial-organizational psychology from Clemson University with a certificate in occupational health psychology (OHP). In addition, she obtained an educational certificate in policy studies from Clemson University. She is a member of the Society of Industrial Organizational Psychology (SIOP) and has also contributed to several conference presentations and peer-reviewed journal articles on topics of sleep, stress, health, and well-being in organizational settings.

Dr. Summer Rebensky, Senior Scientist, Aptima, Inc. has 10 years of expertise focusing on human performance, cognition, and training in emerging systems. In her role at Aptima, Inc. she serves as the portfolio lead for Air Force training, learning, and readiness technologies. She specializes in leading efforts to develop, test, and implement augmented reality (AR), virtual reality (VR), extended reality (XR), and modern training solutions to improve and measure human performance. Dr. Rebensky has previous experience as a postdoctoral research fellow as a part of the Air Force Research Laboratory's (AFRL's) Gaming Research Integration for Learning Laboratory (GRILL) conducting research on interface designs for novel civilian and DoD aviation use cases as well as human-agent teaming designs. Her other experience includes leading Florida Tech's ATLAS lab research efforts for the Air Force, Navy, and FAA; human factors assessments of aircraft with Northrop Grumman; and developing DoD courseware with Raytheon. Her research experience involves leveraging cutting-edge technology to optimize human performance in training and operations, individualized and adaptive training, human-agent teaming, and trust in AI. Dr. Rebensky received her BA in psychology, MS in aviation human factors, and PhD in aviation sciences with a focus in human factors from Florida Tech.

Thomas Dubai, System Engineer, Serco, is a former US Navy sailor whose primary interest was in Ship Self-Defense training and has worked with multiple platforms and training systems to provide tactical training for Amphibious groups. After his service, he transitioned to NCTE project and worked on the Surface Ship Integration team and lead the NCTE Training Baseline requirements against new weapons systems in testing environments. Currently, he is working with the CAF DTC and transitioning operational environment to virtual training as well as integrating new technologies into the DMO program such as JSE. Tom Dubai holds a BS in Physiology from Virginia Commonwealth University.

Nikola Cardenas, System Engineer, Serco, brings eight years of experience supporting Department of Defense network operations, both in uniform and as a civilian. He began his career in the United States Air Force, where he served in a variety of communications and network engineering roles across multiple assignments, including enterprise IT support, cybersecurity enforcement, and large-scale network administration. During his Air Force service, Nikola supported global C4I and intelligence missions, managed secure network infrastructure across multiple enclaves, led communications support teams for thousands of users, and enforced cybersecurity compliance in complex, high-tempo environments. His experience includes both tactical and enterprise-level operations, giving him a strong foundation in mission-driven, secure IT practices. At Serco, he supports distributed training and simulation programs by maintaining secure, resilient network environments. His work includes designing virtualized infrastructure, managing segmentation and access controls, and ensuring systems align with DoD cybersecurity standards. Nikola holds a bachelor's degree in network operations and security and maintains a range of professional certifications in networking, cybersecurity, and cloud administration. He brings a practical, operations-focused approach to solving complex technical challenges in support of national defense readiness and training initiatives.

Generative AI Models, Agents and Tools for Multimodal Training Scenario Generation and Mission Planning

Will Dupree, Svitlana Volkova, Hsien-Te Kao, Grant Engberson, Miles Markey, Gabe Ganberg, Alexxa Bessey, Summer Rebensky
Aptima, Inc.
Woburn, Massachusetts
{wdupree, svolkova, hkao, gengberson, mmarkey, gganberg, abessey, srebensky}@aptima.com

Thomas Dubai
Nikola Cardenas
Combat Air Force (CAF) Distributed Training Center (DTC)
Langley AFB Hampton VA
{Thomas.dubai.ctr, Nikola.cardenas.1.ctr}@us.af.mil

INTRODUCTION

The modern battlefield is rapidly evolving, necessitating advanced training methods for warfighters to prepare for emerging threats (Joint Chiefs of Staff, 2018; Department of Defense 2020). A crucial aspect of this training, particularly for operators and operators in the Air Force, involves exposure to realistic missions and scenarios through virtual simulations (Zook et. al., 2012). These simulations allow operators to train against Computer Generated Forces (CGFs) in a low-risk environment, reducing costs while maintaining the complexity needed for skill acquisition and task repetition. However, creating CGF scenarios remains challenging. The design, planning, and orchestration of large-scale wargaming scenarios can be prohibitively expensive and time-consuming, often requiring multiple subject matter experts (SMEs) to invest hundreds of hours (Nielson and Kratiak, 2021; Fletcher 2009). This does not include the technical effort needed to transcribe scenario documentation, such as planning sheets and simulated operation briefings, into CGF software platforms (e.g., Next Generation Threat System, ASCOT 7). Additionally, SMEs must ensure that mission planning information meets the requirements for developing adversary tactics, techniques, and procedures (TTPs), resulting in a process that, while necessary for accuracy and relevance, is often brittle.

Generative AI models and agents (Anthropic 2023; Open AI 2021; Google 2023) offer transformative potential for revolutionizing CGF scenario creation by dramatically reducing the time and expertise required for mission design. These AI systems could automatically generate diverse, realistic training scenarios by learning from existing mission documentation and SME inputs, then producing variations that maintain tactical validity while introducing novel challenges (Aptima 2023). AI agents (Luo et al., 2023) could serve as intelligent assistants that translate natural language scenario descriptions directly into CGF software formats, eliminating the manual transcription bottleneck and reducing hundreds of hours of work to minutes. Furthermore, generative AI models could continuously evolve adversary TTPs based on real-world intelligence and emerging threats, ensuring training scenarios remain current and adaptable without substantial human intervention. This adoption allows for efficient decision support, allowing the human to guide artifact creation to efficiently scale novel scenario generation. By combining model's understanding of military doctrine with specialized training on tactical scenarios (Caballero and Jenkins, 2023), such AI systems could provide access to high-quality training simulations, allowing even smaller units without dedicated SME teams to create and customize complex, multi-domain training exercises. Scenario generation via the presented novel compound AI approach (Volkova et al., 2024; Zacharia et al., 2024) is a complex task, involving multiple components large language models (LLMs), vision language models (VLMs), agents and tools that work together to produce outputs useful to planners and wargamers. In this paper we introduce a compound AI approach that utilizes training and tactical documents at scale to retrieve, summarize and generate training scenarios across modalities including text, xml encoding and visuals (see Figure 1).

Table 1. Example operator prompts to enable the proposed multimodal scenario generation approach.

Extract Structured Mission Knowledge	Create Novel Scenarios
XML Prompt: Using this scenario file please break it into chunks (e.g., 50 lines per segment) and summarize that chunk *Image Prompt: For this NGTS screenshot please extract number of aircraft* *Text Prompt: Given this tactics document please extract "attack formations" and describe them*	*XML: Using this chunk that describes XML code for an F-16 make a new example for an F-35* *Image and Text: Given details from this image and dogfight posturing, please make an XML with two attacking red forces in Montana airspace*

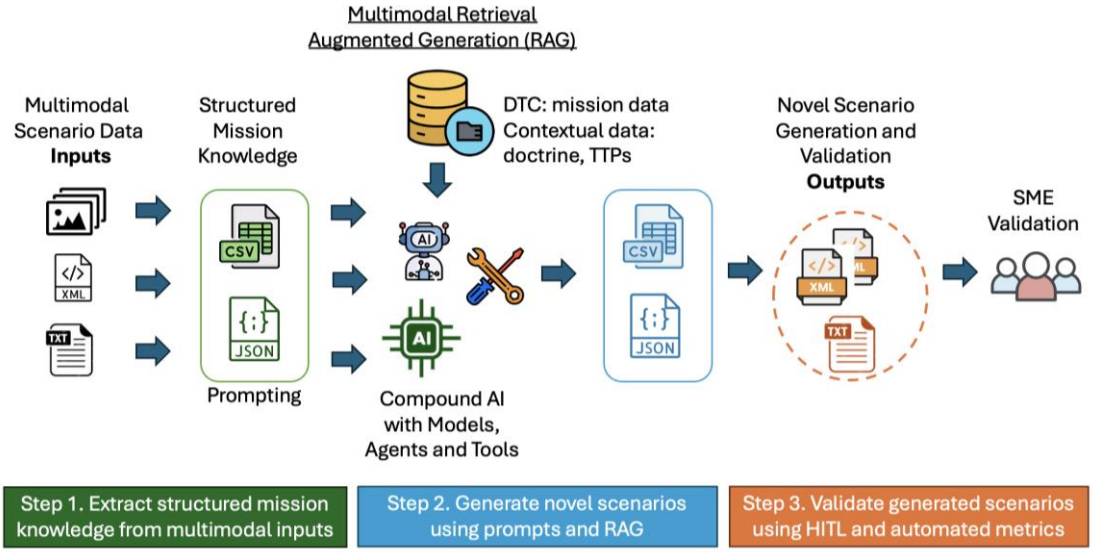

Figure 1. A compound AI approach to retrieve, synthesize and generate new training scenarios. Through synthesis of multimodal inputs combined with structured knowledge representations and RAG-based retrieval, ForceGen extracts mission-relevant knowledge necessary for the creation of new scenarios; structured outputs in json or xml could be further tailored to operational needs e.g., CGF initial conditions or instructions for operators.

At scale, ForceGen processes images, text, and source code into structured representations (step 1) readily available for the compound AI generation components downstream e.g., LLMs with RAG, agents and agentic AI workflows. Then, the system components use such multimodal data via LLM or VLM models or agents to generate new scenarios (step 2). Finally, generated scenarios are validated through automated doctrinal compliance checks and XML schema verification, and human-in-the-loop validation ensuring tactical validity and CGF platform compatibility before deployment to operational training environments (step 3). Note, content generation from the compound AI system is similar to other chatbot like applications, where a user provides a query and then models, agents or tools use aligned context to make new scenario content. In the next two sections we further detail how extraction occurs per modality and provide a summary of how the compound-AI system operates.

This modular approach offers (1) improved mission-relevant knowledge extraction from unstructured and semi-structured multimodal inputs (e.g., images, text, xml etc.), (2) complementary context retrieval based on the most relevant documents to the query using RAG, and (3) multimodal input and output format generation tailored to user needs and workflows for mission planning and SME validation. Below, we detail our methodology, experiments and results for structured representation learning and mission-relevant knowledge extraction to enable downstream training scenario generation from multimodal inputs. We demonstrate the system's capabilities across text, visual, and XML modalities. The paper is organized as follows: in the background section, we discuss the state-of-the-art approaches of AI/ML for training content generation. The methodology section describes large-scale operational dataset from DTC, details the multimodal input processing approach to set up multimodal RAG, and presents the system's backend architecture. The experiments and results section demonstrates mission-relevant knowledge extraction from visual content using a 5-step workflow and discusses structured knowledge extraction from XML scenario files; finally, it presents demonstrations of the approach's XML-based scenario generation capabilities. We conclude with discussions on limitations, address operational implications and provide future.

BACKGROUND

AI/ML Approaches for Training Scenario Generation: Recent advances in AI/ML offer the potential for more rapid wargaming plan creation, scenario generation, and course of action (COA) planning (Goecks, et al., 2024, Sottilare, et al., 2025, Schatz, et. al., 2024). Litvinas (2024) applied Hidden Markov Models (HMMs) for predictive military planning, using probabilistic state transitions to forecast enemy courses of action. Lebanoff et al. (2023) evaluated behavior cloning methods for agent development, replicating expert behaviors through imitation learning. While these approaches address specific aspects: HMMs for predictive planning and behavior cloning for agent development—they remain limited compared to modern multimodal generative AI systems. The shift from these

statistical methods to compound AI architectures represents a paradigm change, enabling end-to-end scenario generation that integrates doctrine, visual tactics, and executable code in unified frameworks.

This paper presents a novel approach that leverages LLMs (Minaee, et al. 2024) and VLMs (Li, et al., 2025) in an end-to-end pipeline for mission-relevant knowledge extraction and training scenario generation. Building on our previous work in compound-AI approaches for training (Volkova et al., 2024; Rebensky et. al., 2024), as well as structured information extraction, LLMs with RAG (Aptima 2023; Volkova et al., 2025a) and agentic AI workflows for decision advantage (Volkova et al., 2025b; Klinefilter et al., 2025), we now apply them to extract mission-relevant knowledge from operational datasets and generate scenarios and COA plans. RAG comes in many forms, such as semantic based similarity (Lewis et. al., 2020, Sachan et al., 2021) or structured information via knowledge graphs (KG) (Microsoft, 2023, Sarmah, et al., 2024. This approach uses context (e.g., background TTPs and past scenario descriptions) to ground open and closed LLMs (Touvron et al., 2023) and VLMs in its response to a user query as described using prompt examples in Table 1. Other approaches may use rule-based expert systems, finite state machines, and scripted behavior trees**,** which are beneficial for deterministic scenario control, predictable agent behaviors, and transparent decision logic**,** but require extensive manual coding, domain-specific programming expertise, and inflexible architectures that cannot adapt to novel tactical situations**.**

While text-based information provides details useful at inference time, operators do not perform scenario planning in text formats only. Scenario generation capabilities must also make use of image-based scenario representations and xml code for complete and accurate generations. As discussed by Cheng (2025) and Roberts (2024), AI research community is improving actionable data that can be drawn using generative AI models from both text and visuals. VLMs has been very sparsely leveraged for training applications. However, general purpose models have been applied to downstream tasks like image captioning (Open AI, 2021) to enhance and validate the information perceived by VLMs. As applied to scenario generation, artifacts and materials being used in planning involve Word documents, Excel spreadsheets and PowerPoint (PPT) slides that include images of platform posturing (e.g., latitude and longitude locations), battlespace and area descriptions (e.g., target locations, no-fly zone markings), as well as force movements and strategic routes. The range of information encoded in images ranges from simple backgrounds to complex heterogenous images with many markings, assets (ground or air), and noisy map backgrounds. This requires instructors and planners to navigate large briefings to prepare wargamers and CGF software for successful simulate mission execution (Volkova et al., 2025; Sottilare et al., 2025).

METHODOLOGY

Operational Dataset Overview: To generate scenarios and COA details, we focused on several sources of information. The first is text-based documents like that which are seen in the CGF mission planning phase e.g., the TTP and PPT documents. The second form of data is based on software source code and machine-based text that underpins virtual wargaming within the CGF software. Note, the approach applies a code-extraction framework that is agnostic of which CGF files a user provides. This extends future domains and mission types that this approach could be applied to. Finally, the last form of data is the combination of visuals and text extracted from multimodal documents (slides, operational maps, and annotated imagery) that capture spatial relationships, tactical positioning, and temporal sequences critical for scenario fidelity. These multimodal data inputs are important for LLM-based applications for scenario generation to synthesize complex operational contexts that text alone cannot adequately represent, ensuring generated scenarios accurately reflect the visual-spatial reasoning inherent in military planning and execution (Sottilare et al., 2025).

Table 2. DTC corpus statistics ingested into the compound AI system's multimodal RAG backend.

	Number of files	Storage Volume (GB)
pptx	3,470	26.0
xlsx	5,977	6.8
xlsm	382	4.1
pdf	1,408	2.4

Due to mission relevance, we are unable to share real-world examples of training scenario. It resembles planning materials relevant to the Distributed Training Center (DTC), within the Combat Air Force (CAF) Distributed Mission Operations (DMO) and consists of briefings and planning files for large scale wargaming exercises. However, in Table 2 we present data statistics for text-based documents with TTPs for fighter operator-based missions, Excel/PPT sheets

and training outcome briefings. We leverage these documents to develop our compound-AI approach (Volkova et al., 2024) for mission-relevant knowledge extraction and multimodal scenario generation.

To augment the DTC dataset and ensure generation validation, we simulated additional XML files that share complexity and characteristics to SOTA CGF packages (e.g., NGTS, NICE, ASCOT 7). The initial goal for scenario generation is to focus on the *first step in the CGF operator training process, transitioning planning documents into initial condition within the CGF*. This includes scripting what platforms, loadouts, and locations will be used. Depending on the CGF there is also behavior models used for the constructive forces (e.g., the simulated red forces) or CGF-specific assets. To this end, we created synthetic XML files that describe initial conditions for a CGF scenario and performed mission-relevant knowledge extraction that can be harnessed for generating new XML in the same format as the synthetic files. As presented in Figure 1, the approach first focuses on using all available multimodal and unimodal data to extracting key insights relevant to scenario generation. This occurs as *creating structured outputs*, e.g., extracting scenario details from an image and putting them into a table to verify key details.

ForceGen Backend Architecture: The approach implements a novel compound AI architecture that orchestrates LLMs and VLMs with Retrieval Augmented Generation, that could be further augmented with AI agents and specialized tools to enable automated generation of large-scale military scenarios and COA details. The system's RAG implementation employs a hybrid multimodal search strategy using LlamaIndex, which indexes text-based doctrinal materials (TTP documents, PPT briefings), code-based technical information from CGF software files, and visual content extracted from operational planning materials. For text and code embeddings, it utilizes sentence-transformers/all-mpnet-base-v2, which provides technical military terminology and programming constructs compared to lighter-weight models. Visual content from PowerPoint slides, tactical maps, and annotated imagery is summarized using a pre-trained vision language model (VLM) to enable, enabling cross-modal retrieval between textual scenario descriptions and visual tactical representations. These multimodal embeddings are stored in OpenSearch vector database for rapid retrieval performance across documents and images. The system implements a three-stage retrieval process: initial semantic search with cosine similarity matching, re-ranking using cross-encoder models, and final filtering based on doctrinal compliance scores. This multi-source knowledge integration ensures that generated scenarios are both tactically valid according to military doctrine and technically implementable within CGF platforms, while visual grounding ensures spatial accuracy for entity positioning and movement patterns. Built on AWS infrastructure using r5.8xlarge instances for compute and S3 for distributed storage of multimodal assets, this architecture would allow reduced scenario creation time from hundreds of hours to minutes while maintaining the complex spatial-temporal relationships critical for realistic scenarios.

Extracting Structured Mission-Relevant Knowledge from Multimodal Scenario Data

Popular foundation models have been capable of generating text in the form of structured data (e.g., JSON) for years (OpenAI, 2023), as well as more recently explored for improved performance of styled outputs (Lu, et al., 2025). This structured output format is popular when using LLMs to create machine-to-machine responses, assisting software pipelines that invoke LLMs in intermediary steps. By representing the output of text-to-text with defined schemas we give LLMs the power to use new toolsets and extend the tasks they can complete (Gupta & Kembhavi, 2022). The approach leverages structured outputs from LLMs and VLMs to improve representation of mission-relevant knowledge for future novel scenario generation. We do this by focusing on the text, image, and XML files available as presented in Figure 2, and instructing LLMs to parse out relevant details from the context presented.

Text: The first step is to take long documents and turn them into smaller segments, often called “chunks”, and save these chunks for later retrieval when needed for a task. For applications of RAG, chunks are used by finding similar text segments to user queries and giving them as context to provide improved answers without hallucination. We can also task the LLM with named entity extraction (NER), a common NLP task, where key items are identified by name. Both chunks and entity outputs can be templated or formatted into tables by the LLM. Structuring the output improves the focus of the LLM on key details when it is time to create new content.

Image: As discussed, VLMs have been shown useful in creating text summaries and captions when presented with visual data (Cheng, et al., 2025). When applied to scenario generation, this includes describing the image at different levels of complexity and adding the summarized details to a table. Similar to human analysis, by addressing the image extraction in steps we provide the VLM with a framework to create clear insight about homogenous (e.g., sets of images with the same style) and heterogenous (e.g., complex scenario planning images that vary widely based on mission need) images. The current approach is to first identify basic spatial elements, such as number of aircraft and positions. Next, we identify team (e.g. friendly vs. adversary based on blue/red descriptions) and intent. As we scale, past extracted context can be leveraged alongside the image to improve the analysis. Finally, we attempt representing

strategy and training objectives based on the previously retrieved and summarize knowledge by the VLM. This gives a comprehensive view of the mission from images only, ensuring natural points of refinement for improved structured representation and image analysis. We apply a multi-step approach through a series of prompts, each working towards one of the sub-task goals of turning visual representations into structured knowledge tables for future use. As a result, text summaries from tables or captions created from the images can be used to identify additional relevant images of importance when planning new scenarios or COA, enhancing creation prompts with extracted information not contained in text alone.

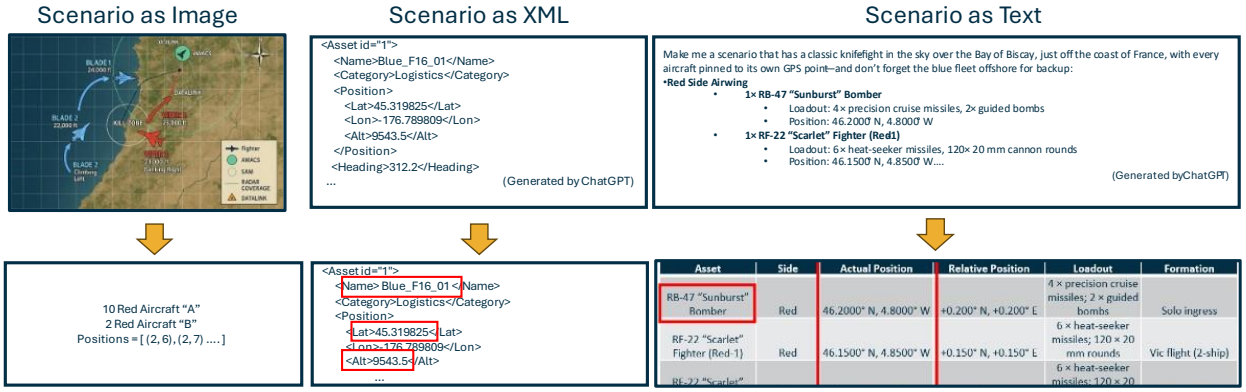

Figure 2. Example scenarios inputs encoded as text, images or xml and the corresponding outputs – mission relevant knowledge extracted e.g., named entities, object counts, position coordinates (OpenAI [o4-mini-high], 2025).

XML: Understanding and representing machine code is similar to a translation task using LLMs, where we want to translate between natural text and the text used to give software instructions (e.g., NGTS scenario files). The biggest challenge in representing XML is the need for clear chunking that retains meaning once text is segmented. After chunking, the extraction strategy remains similar to text or image descriptions given above. Each code chunk can be summarized in plain English, as well as NER-like concepts extracted for hard-coded (e.g. strictly named) elements and identifiers. Other attributes that define code structure and formatting can also be extracted, such as named levels of nested components or file names. Such chucking strategy allows quickly creating accurate new asset representations, tailored to a specific CGF based on provided files as context.

EXPERIMENTS AND RESULTS

Multimodal Data Processing and Encoding for Retrieval Augmented Scenario Generation

Retrieving and synthesizing mission-relevant knowledge from multimodal media, such as power point slides that contain a mixture of text, embedded images and tables with associated captions, presents unique challenges not adequately handled by current off the shelf generative AI solutions. Text extraction alone from these files fails to sufficiently capture their embedded information since these file types tend to rely heavily on their embedded media, meaning that much of their intended message is not explicitly delivered in text. Additionally, since these file types are intended to be presented and discussed visually, traditional image extraction and summarization, which typically involves first extracting embedded images and tables from the files then generating summarizations for these extracted components using VLMs, is also insufficient because it fails to capture all of the relevant context that these embedded files are presented alongside, such as captions, text, or associated figures presented in the same slide. Instead, we explored applications for multimodal data processing as described in Figure 3 that incorporates image summarization of full context items

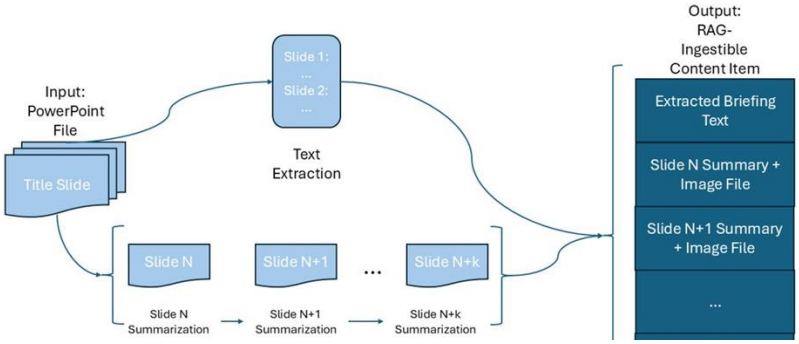

Figure 3. Multimodal data processing pipeline for RAG-based generation for a sample mission briefing slide.

e.g., PowerPoint slides. This method preserves the original intent of the multimodal media by providing the VLM with the same view of the media that human users would see, removing unnecessary media parsing that would limit its context. As seen in Figure 3, multimodal media extraction consists of two steps. The first is a straightforward text extraction pipeline that extracts all text from the media and stores it in the resulting content item in the backend database. This is done to aid in retrieval, because while the text in these files is not sufficient for summarization or understanding these multimodal files, it is still necessary for the vector similarity algorithm the system uses when searching for relevant documents from the corpus of documents that it is given. Context aware summarization is done in parallel to text extraction and adds a VLM-generated summary of the file to the resulting content item. This summary is intended to capture visual context of text, embedded images, tables, diagrams etc. that is not processed during text extraction. Such context-aware summarization consists of sampling *k* slides from each presentation and running *k* successive VLM summarization prompts, with each prompt including a summary of all previously generated prompts. This not only gives the resulting content item a visually aware summarization but enables these summaries to build upon each other just as a human reader would interpret files whereas simple text extraction would miss all this visual context and media extraction would miss larger scale context regarding how extracted media is intended to be interpreted in conjunction with other media it is presented alongside.

Mission-Relevant Knowledge Extraction from Visual Scenario Content

Extracting mission-relevant knowledge from complex visual scenarios as demonstrated in Figure 4 involves multiple challenges that test the limits of current AI/ML capabilities. VLMs must interpret both static elements like objects and symbols and dynamic cues such as arrows, trajectories, and spatial arrangements that indicate movement, intent, or tactical relationships. These directional and action-based signals are central to understanding the image. The complexity increases with layered compositions where background textures, overlapping icons, color gradients, and spatial context each carry functional or symbolic information. This requires VLMs to go beyond surface-level detection and engage in deeper contextual reasoning. Additionally, VLMs must be ready to extract both obvious and subtle signals without relying on predefined categories. Multimodal content adds to the difficulty because tactical visuals often combine charts, text, numbers, and metaphors in one message. VLMs must link these diverse forms to build a coherent understanding. Finally, open-form content resists simple classification and requires adaptive, context-aware extraction methods that can evolve with the content. Together, these challenges demand flexible compound AI solutions capable of nuanced visual understanding. To meet these challenges, we propose a step-by-step framework for extracting mission-relevant knowledge from visual content that guides VLMs through an adaptive understanding process, as shown in Figure 5. First, the system classifies the image as static or dynamic by identifying cues like arrows or motion trails. This classification shapes the approach for interpretation. Next, the system summarizes the image by extracting a few high-level points that capture the main message, reducing complexity while maintaining clarity. Then, the system identifies relevant entities, which are the key objects or components, and filters out irrelevant elements. For each entity, it selects essential attributes needed to understand its role, adjusting these based on context. Finally, the system compiles structured data linking entities to their attributes, creating an interpretable representation for reasoning or further downstream tasks e.g., new scenario generation and planning. This multi-step approach allows the system to handle complexity without losing semantic depth by processing information in context and breaking it into manageable units. It improves accuracy, efficiency, and generalizability across domains.

Figure 4. Synthetically generated battle map aircraft combat generated by GPT-o3 to show image inputs to models.

Applying the system's knowledge extraction pipeline from visual content with a GPT-o3 model generated battle map for aircraft combat shows its effectiveness, as demonstrated in Table 3. The system first classifies the image as dynamic by detecting directional elements such as arrows and flight paths that show aircraft maneuvers. It summarizes key information, including the positions and altitudes of two blue fighter jets, 'BLADE 1' and 'BLADE 2,' moving toward the coast at 24,000 and 22,000 feet, with 'BLADE 2' climbing left. It notes a red fighter jet, 'VIPER 2,' approaching from the east at 23,000 feet and banking right toward a marked 'KILL ZONE' where engagement may occur. The system also identifies an inland AWACS platform that provides radar and datalink support. It then identifies relevant entities including the jets with labels and altitudes, the kill zone, AWACS, radar coverage circles, datalink lines, and a compass rose. For each, it extracts necessary attributes such as motion cues, textual labels, color-coded

team affiliation, and static elements like radar and communication lines. This detailed extraction creates a rich, structured representation of the tactical environment that integrates spatial and semantic information. Breaking the image into these parts ensures all dynamic, static, and multimodal signals are captured effectively, supporting decision-making and situational awareness in combat scenarios.

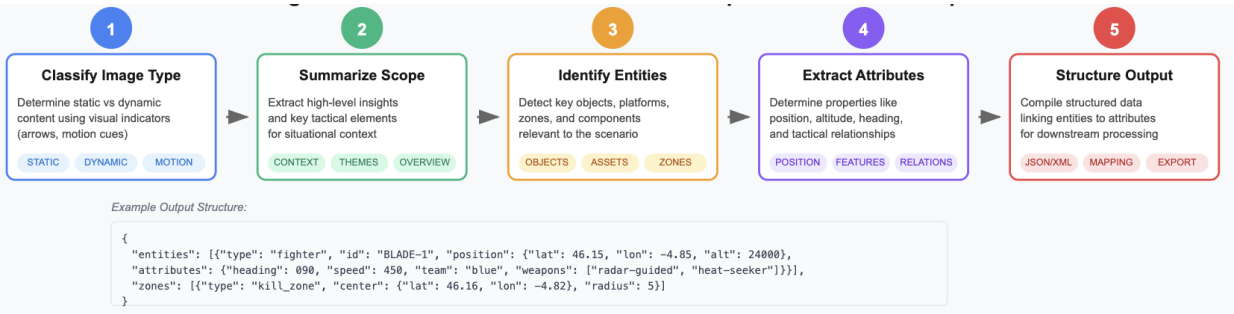

Figure 5. Mission-relevant knowledge extraction workflow from visual content using VLMs.

Table 3. Structured mission-relevant knowledge extracted using the AI system from synthetic tactical battle map.

Category	Entity	Key Attributes	Description
Blue Aircraft	BLADE 1	Position: 24,000 ft altitude Motion: Maneuvering toward coast Color: Blue	Blue fighter jet labeled "BLADE 1" at 24,000 ft, approaching coast with specific motion cues
Blue Aircraft	BLADE 2	Position: 22,000 ft altitude Motion: Climbing left Color: Blue	Blue fighter jet labeled "BLADE 2" at 22,000 ft, climbing left with curved motion indicators
Red Aircraft	VIPER 2	Position: 23,000 ft altitude Motion: Banking right toward KILL ZONE Color: Red	Red fighter jet labeled "VIPER 2" at 23,000 ft, banking right, approaching from east toward engagement area
Engagement Zone	KILL ZONE	Type: Static element Purpose: Potential engagement area Color: Red/Orange	Critical area where blue and red jets intersect, indicating potential engagement point
Support Systems	AWACS	Type: Airborne Warning and Control System Position: Positioned inland Function: Radar coverage and datalink support	Provides radar coverage and datalink support to fighters
Infrastructure	Radar Coverage	Type: Detection circles Function: Indicates radar coverage area	Circles showing radar detection and tracking capabilities
Infrastructure	Datalink Lines	Type: Communication lines Function: Data sharing connections	Lines connecting AWACS to fighters for communication and data sharing
Navigation	Compass Rose	Type: Static reference Function: Cardinal directions	Provides navigational context for the tactical scenario

Mission-Relevant Knowledge Extraction from XML Scenario Content

XML scenario files serve as the bridge between human tactical intent and machine-executable simulations, containing both human-readable metadata (aircraft callsigns, mission objectives) and machine-specific configurations (spawn coordinates, behavior trees, weapon loadouts) that must be parsed with semantic awareness to enable meaningful manipulation and generation. Naïve chunking methods fragment entity configurations across multiple segments, breaking critical relationships between aircraft, weapons, and positioning data. For example, an F-16's configuration might be split such that its weapon loadout appears in one chunk while its mission parameters appear in another, making it impossible to understand or replicate the complete tactical setup. Preliminary experiments feeding large XML files directly to LLMs resulted in hallucinated fields and malformed code due to attention dilution across complex nested structures.

To address these challenges, we developed a three-stage recursive semantic decomposition process that preserves tactical relationships while enabling scalable processing while chunking. First, an LLM analyzes XML element tree structure to identify "atomic chunks"—logically inseparable code blocks that preserve complete entity relationships (e.g., an aircraft with its full configuration). Next each chunk is classified as either operational data (meaningful for scenario execution) or metadata (boilerplate configuration), then summarized in natural language and stored in a vector

database with preserved structural relationships. Finally, user requests trigger semantic search through the hierarchy, identifying relevant code segments for modification or replication. The performance evaluation across 11 scenario files (5 synthetic, 6 operational) demonstrates robust performance (see Table 4). Perfect boundary detection (e.g., isolating desired sub-elements of XML chunks) validates the recursive decomposition approach. Classification challenges in operational data (84% precision vs 100% synthetic) highlight the need for few-shot learning approaches.

Table 4. Experimental results of boundary detection and classification of nodes.

		Boundary Detection	Chunk Classification Metrics		Characters/Tokens		
Source	Level	Accuracy	Precision	Recall	Characters	Tokens	Char./Token
Synthetic Data	0	1.0			4,331	1,423	3.04
	1	1.0	1.0	1.0	432	142	3.04
	2		1.0	1.0	85	25	3.40
Real Data	0	1.0			254,401	52,453	4.85
	1	1.0	0.50	1.0	12,595	2,594	4.86
	2		0.84	0.74	970	195	4.97

This semantic chunking capability enables natural language manipulation of CGF scenarios without requiring SMEs to understand syntax or XML structure. Operators can request modifications like *"move the red forces 50 nautical miles north"* and the system automatically identifies and modifies the appropriate spawn location parameters—transforming a previously manual, error-prone process into an automated, validated operation. The structured XML representations feed directly into the multimodal RAG backend, where they can be retrieved alongside relevant doctrinal text and tactical imagery. This integration enables the system to generate new scenarios that maintain both tactical validity (from doctrinal sources) and technical executability (from XML templates), addressing the complete workflow from mission concept to executable simulation.

Training Scenario Generation and Initial Validation

Next, we explored the ability to generate new scenarios in an XML format using the same simulated data above. It consists of five XML files, all structured to represent wargaming scenarios that include multiple assets (e.g., fighter aircraft, tanks, aircraft carriers) and the weapon systems they may have. As seen in ***Figure 6*** the XML file describes initial conditions for the scenario including position and entity metadata. Given the findings from the semantic chunking above, we take a simple approach of length-based cut points based on character count and XML element tree structure. For the simulated files this includes natural break points at the "Asset", "Position", and "Weapons" sections, ensuring that context rich information stays intact for future context examples. It will be the case for advanced CGFs that the source code is not as simple, and additional consideration will need to be taken to ensure the software's initial condition scenario files are chunked appropriately. We include four items of context per chunk to ground scenario generation. These items are: (1) the raw XML chunk that contains the source code from the file, (2) a summary of the raw code in human-readable text, (3) a copy of the files XML tree structure and file name, and (4) numerical embeddings of the summary text. The produced embeddings can be used to align new queries and requests for scenarios with examples from the XML corpus. By including the additional file metadata, we can better ground the LLM in producing relevant scenario text that matches the XML structure, avoiding hallucinations. Additionally, it lays the foundation for context awareness from co-mingled scenario files from different CGFs. Such a capability would allow simple translations or augmentations between systems, such as moving from NGTS to ASCOT 7.

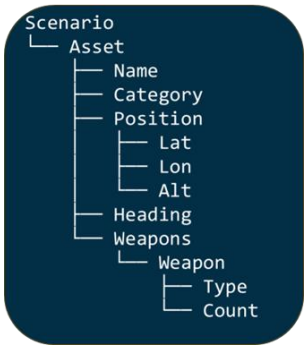

Figure 6. Simulated wargaming scenario XML file structure.

Figure 7. Example scenario generation output based on simulated scenario prompts.

We apply this process to the small set of simulated CGF files and visualize the output via Python and Plotly. The input is a text-based description of a scenario we would like, with descriptions of the aircraft, support assets, and weapons to populate the file with. This can all be passed in with a natural language prompt (also currently simulated), where we focus on a fictional engagement in international waters off the coast of France. Results of the approach's XML output can be seen in Figure 7. We can see the location of the aircraft and vessels is created successfully, where we have a red bomber, 2 red aircraft, 2 blue aircraft, and blue support. While the simulated data is not as complex as NGTS source files, they provide a strong case for CGF agnostic approaches to software scenario generation.

Prompt: "Make a scenario that has a classic knife-fight in the sky over the Bay of Biscay, just off the coast of France, with every aircraft pinned to its own GPS point—and don't forget the blue fleet offshore for backup:
Red Airwing – RB-47 "Sunburst" Bomber (46.2000° N, 4.8000° W): 4 × precision cruise missiles; 2 × guided bombs
Red Airwing – RF-22 "Scarlet" Fighter (Red-1) (46.1500° N, 4.8500° W): 6 × heat-seeker missiles; 120 × 20 mm rounds
Red Airwing – RF-22 "Scarlet" Fighter (Red-2) (46.1800° N, 4.7800° W): 6 × heat-seeker missiles; 120 × 20 mm rounds
Blue Airwing – BF-9 "Azure" Fighter (Blue-1) (46.1600° N, 4.8200° W): 8 × radar-guided missiles; 100 × 25 mm rounds
Blue Airwing – BF-9 "Azure" Fighter (Blue-2) (46.1700° N, 4.7700° W): 8 × radar-guided missiles; 100 × 25 mm rounds
Blue Naval Support – Destroyer "Wavebreaker" (46.1000° N, 4.9000° W): 8 × SAMs; 16 × anti-ship missiles; 1 × CIWS
Blue Naval Support – Frigate "Horizon" (46.0500° N, 4.9500° W): 4 × ASROC torpedoes; 6 × SAMs; 76 mm deck gun"
(OpenAI [o4-mini-high], 2025)

Extending Beyond Generative AI Models to Include Agents and Tools

Building upon the compound AI architectures for military training ecosystems (Volkova et al., 2024), the evolution toward agentic AI workflows represents a natural progression in automated scenario generation capabilities. Future implementations will incorporate specialized agents working in concert: text-only and multimodal *Scenario Planning Agents* that orchestrate the overall workflow by interpreting natural language queries and visual tactical representations, a *Doctrine Retrieval Agent* that extracts relevant information from TTP documents using semantic search across indexed military doctrine, a *XML Analysis Agent* with associated tools that process CGF software source code through domain-agnostic extraction frameworks to understand technical constraints, and a *COA Synthesis Agent* that integrates doctrinal knowledge with technical limitations to generate executable scenarios. Working in concert with SME-driven qualitative evaluation performed to date, the compound AI system will implement *Scenario Evaluation Agents and Tools* that assess generated scenarios against doctrinal compliance, tactical validity, scenario complexity, competency breakdowns, anticipated student performance, and technical feasibility metrics, providing real-time feedback to refine generation parameters. These agents will be augmented by knowledge graph-enhanced RAG (GraphRAG) to maintain semantic relationships between military concepts and TTPs, enabling more nuanced scenario generation. Through scaled deployment with DTC data, these novel agentic workflows will continuously learn from SME feedback and operational outcomes, creating a self-improving system that reduces scenario generation time while increasing tactical realism and training effectiveness

LIMITATIONS

While our work to date demonstrates promising capabilities in multimodal scenario generation, several limitations merit consideration (Rivera et al., 2024). First, the system's performance is inherently constrained by the quality and diversity of the training corpus; scenarios generated may exhibit gaps present in historical DTC documentation and existing TTPs. Second, the current implementation relies on simulated XML files rather than actual CGF software formats, potentially limiting immediate deployment without additional validation and adaptation to specific platforms like NGTS or ASCOT 7. Third, the multimodal processing pipeline, particularly the VLM-based image analysis, may struggle with highly complex tactical visualizations containing dense overlapping symbols, degraded image quality, or non-standard notation systems used by different units. These could be mitigated by fine-tuning models on operational datasets. Fourth, the system currently lacks mechanisms for real-time SME validation and feedback integration, which could lead to tactically questionable scenarios passing initial generation filters. These could be addressed by our team's recent work on simulating operation-agent interactions, benchmarking and evaluation (Volkova et al., 2025a; Volkova et al., 2025b; Lamparth et al., 2024) Notably, the current evaluation relies primarily on qualitative SME assessments without comprehensive quantitative validation metrics or standardized benchmarks for measuring scenario fidelity, tactical validity, or training effectiveness across large-scale deployments. Finally, the data currently used for testing and evaluation of the system is at the unclassified level. For true assessment of mission relevance and benefit, the approach would need to consider scenario, TTP, and vehicle information at classified levels. This will ensure generated scenarios meet mission standards and challenges warfighters face in their day-to-day. Future iterations must address these limitations through expanded training data diversity, direct CGF platform

integration, enhanced image processing capabilities, and automated validation frameworks that incorporate continuous SME feedback loops.

OPERATIONAL IMPLICATIONS

The advancements in AI capability represents a paradigm shift in Combat Air Force readiness generation, directly addressing the critical capability gap between adversary force modernization rates and scenario development timelines. By reducing scenario creation from 200+ SME hours to 10-minute automated generation cycles, it could alter the operational calculus for Distributed Mission Operations (DMO) training throughput. This 1,200x acceleration in scenario production velocity enables daily refreshment of adversary TTPs based on real-world intelligence feeds, ensuring US operators train against threat representations that mirror current operational realities rather than outdated doctrinal templates. Further, it provides access to high-quality, dynamic scenario generation across all training centers and units – regardless of size and access to resources. This allows for more cutting-edge training to occur not just within distributed training offered by training centers but also within local training at the unit level. The system's CGF-agnostic XML translation capability creates unprecedented interoperability across NGTS, ASCOT 7, and emerging JSE platforms, eliminating the technical silos that currently fragment our training enterprise. Most critically, the multimodal RAG architecture and integrated models, agents and tools preserves and propagates hard-won tactical knowledge from senior weapons officers and mission commanders, transforming ephemeral SME expertise into persistent, queryable operational intelligence. There has been a long-standing desire for tactically relevant, adaptive, and novel training capabilities that use tomorrow's tactics today. The limitation has long been the information flow and standup of new scenarios. The benefit of the multimodal generation approach is that AI benefits can be applied to legacy systems now. However, as noted in the limitations, advancements in these areas requires more open architecture of existing and emerging systems. Government reference architectures will enable the government to define inputs and outputs of different solutions allowing a more streamlined plug and play of not only commercially available AI solutions, but also standardized data formats making it easier to train up model-based solutions without being tailored to a specific simulation platform system. Advancements have been made in efforts such as Simulator Common Architecture Requirements Standards (SCARS) and JSE GRID systems to define some of the inputs and outputs and sim connection communication protocols, but similar efforts will be needed for training and scenario data storage to enable AI solutions to scale with the architecture design.

CONCLUSIONS

The work demonstrates the transformative potential of compound AI systems for military training scenario generation, reducing creation time from hundreds of hours to minutes while maintaining tactical validity through multimodal knowledge extraction and novel scenario generation. However, challenges remain: LLMs struggle with precise XML generation for CGF platforms and VLMs have limited spatial reasoning capabilities for complex tactical visualizations. Future work will focus fine-tuning vision-language models for military imagery and deploying advanced RAG techniques e.g., GraphRAG jointly with agentic AI workflows. Additionally, we will establish comprehensive test and evaluation framework, starting with simple scenarios before scaling to complex multi-domain engagements, ultimately positioning AI as a critical enabler for next-generation military training that bridges human expertise and automated scenario generation.

ACKNOWLEDGMENTS

Our team would like to thank the Distributed Mission Training Center (DTC) and its leadership, to include Mr. James White, for their invaluable support throughout the development of this paper. We also would like to acknowledge Air Combat Command (ACC), under which the DTC operates, for enabling and supporting this work. The authors have made use of large language models (LLMs) to assist in summarizing or organizing minor details presented in this work. The ideas provided belong to those of the authors and do not reflect or serve as statements or policy of any other US government organization.

REFERENCESu

Anthropic. (2023). Anthropic's Constitutional AI: Claude. Anthropic Blog. https://www.anthropic.com/index/introducing-claude

Aptima. (2023). Instructional system design (ISD) analysis using AI large language models. Technical report. https://www.aptima.com/wp-content/uploads/2024/02/NAUTICAL-White-Paper.pdf

Cheng, K., Song, W., Fan, J., Ma, Z., Sun, Q., Xu, F., Yan, C., Chen, N., Zhang, J., & Chen, J. (2025). CapArena: Benchmarking and Analyzing Detailed Image Captioning in the LLM Era. arXiv preprint arXiv:2503.12329. https://arxiv.org/abs/2503.12329

Department of Defense. (2020). DoD Instruction 1322.35: Military Training. https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/132235p.pdf

Fletcher, J. D. (2009). Education and training technology in the military. Science, 323(5910), 72-75.

Folsom, P., Kovarik, A., Rowe, J., & Lester, J. (2025). Toward automated scenario generation with deep reinforcement learning. In Generalized Intelligent Framework for Tutoring (GIFT) Symposium.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). Retrieval-Augmented Generation for Large Language Models: A survey. arXiv preprint arXiv:2312.10997. https://arxiv.org/abs/2312.10997

Goecks, V. G., & Waytowich, N. (2024). COA-GPT: Generative Pre-trained Transformers for accelerated Course of Action development in military operations. International Conference on Military Communication and Information Systems (ICMCIS)

Goutière, R., Lourdeaux, D., & Lagrue, S. (2023). Automated planning for military airline controller training scenarios. In Proceedings of the 15th International Conference on Agents and Artificial Intelligence (ICAART).

Gupta, T., & Kembhavi, A. (2022). Visual programming: Compositional visual reasoning without training. arXiv preprint arXiv:2211.11559. https://arxiv.org/abs/2211.11559

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., & Qin, B. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. arXiv preprint arXiv:2311.05232. https://arxiv.org/abs/2311.05232

J.-P. Rivera, G. Mukobi, A. Reuel, M. Lamparth, C. Smith and J. Schneider, Escalation Risks from Language Models in Military and Diplomatic Decision-Making, in The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT 24), 2024.

Joint Chiefs of Staff. (2018). Joint Publication 3-0: Joint Operations. https://www.jcs.mil/Portals/36/Documents/Doctrine/pubs/jp3_0ch1.pdf

Keno, H., Pioch, N. J., Guagliano, C., & Chung, T. H. (2024). Simulation-based scenario generation for robust hybrid AI for autonomy [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2409.06608

Lebanoff L, Paul N, Ballinger C, Sherry P, Carpenter G, Newton C. A comparison of behavior cloning methods in developing interactive opposing-force agents. In The International FLAIRS Conference Proceedings 2023 May 8 (Vol. 36).

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Riedel, S. (2020). Retrieval-augmented generation for knowledge-intensive NLP. In Advances in Neural Information Processing Systems (NeurIPS).

Li, Z., Wu, X., Du, H., Nghiem, H., & Shi, G. (2025). A Survey of State of the Art Large Vision Language Models: Alignment, Benchmark, Evaluations and Challenges. arXiv preprint arXiv:2501.02189. https://arxiv.org/abs/2501.02189

Litvinas M. Tip of the Spear: Developing Predictive Military Planning Tools Using Hidden Markov Models. In The International FLAIRS Conference Proceedings 2024 May 13 (Vol. 37).

Lu, Y., Li, H., Cong, X., Zhang, Z., Wu, Y., Lin, Y., Liu, Z., Liu, F., & Sun, M. (2025). Learning to generate structured output with schema reinforcement learning. arXiv preprint arXiv:2502.18878. https://arxiv.org/abs/2502.18878

Luo, J., Zhang, W., Yuan, Y., Zhao, Y., Yang, J., Gu, Y., Wu, B., Chen, B., Qiao, Z., Long, Q., Tu, R., Luo, X., Ju, W., Xiao, Z., Wang, Y., Xiao, M., Liu, C., Yuan, J., Zhang, S., Jin, Y., Zhang, F., Wu, X., Zhao, H., Tao, D., Yu, P. S., & Zhang, M. (2025). Large language model agent: A survey on methodology, applications and challenges. arXiv. https://arxiv.org/abs/2503.21460

M. Lamparth, A. Corso, J. Ganz, O. S. Mastro, J. Schneider and H. Trinkunas, Human vs. Machine: Behavioral Differences between Expert Humans and Language Models in Wargame Simulations, in Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2024.

Microsoft. (2023). GraphRAG: Unlocking LLM discovery on narrative private data. Microsoft Research Blog.

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large language Models: A Survey. arXiv preprint arXiv:2402.06196. https://arxiv.org/abs/2402.06196

Nielson, P. E., & Kratiak, C. (2021). The future of military training: Blending live and virtual in a synthetic training environment. Military Medicine, 186(Supplement_1), 38-45.

OpenAI, DALL-E: Creating images from text, 2021. [Online]. Available: https://openai.com/index/dall-e/

OpenAI. (2023). ChatGPT (o4-mini-high, 2025) [Large language model]. https://chat.openai.com/

Reus, N. d., Schadd, M. P., Maaiveld, T. M., Nieuwenhuis Nyegaard, D. D. N., van den Broek, C. A. A., & van der Waa, J. (2024). Generating explainable military courses of action. In 2024 International Conference on Military Communication and Information Systems (ICMCIS). IEEE.

Roberts, J., Lee, T., Wong, C. H., Yasunaga, M., Mai, Y., & Liang, P. (2024). Image2Struct: Benchmarking Structure Extraction for Vision-Language Models. In Proceedings of the Neural Information Processing Systems (NeurIPS) Conference (2024). https://neurips.cc/virtual/2024/poster/97829

S. Pichai and D. Hassabis, Introducing Gemini: our largest and most capable AI model, Google, 2023. [Online]. Available: https://blog.google/technology/ai/google-gemini-ai/.

Sachan, D. S., Tay, Y., Aralikatte, R., & Yang, Y. (2021). End-to-end training of retriever-reader models with reinforcement learning. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL 2021).

Sarmah, B., Mehta, D., Hall, B., Rao, R., Patel, S., & Pasquali, S. (2024). HybridRAG: Integrating knowledge graphs and vector retrieval augmented generation for efficient information extraction. In Proceedings of the 5th ACM International Conference on AI in Finance (pp. 608–616). ACM. https://doi.org/10.1145/3677052.3698671

Schatz, S., Stodd, J., & Stead, G. (2024). Navigating the generative AI revolution [Tutorial]. Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC).

Sottilare, R. A., McGroarty, C., Ulinski, M., & Osmond, S. (2025). Revolutionizing military training and simulation with multimodal generative artificial intelligence. Proceedings of International Training Technology Exhibition and Conference (ITEC).

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., ... & Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. arXiv preprint arXiv:2302.13971.

Volkova, S., Kao, H.-T., Penafel, L., Nguyen, D., Cohen, M., Lynch, S., McCormack, R., & Cassani, L. (2025). VirTLab-Eval: Human-agent team and digital twin performance evaluation demonstration. In *Proceedings of the Human Factors and Ergonomics Society ASPIRE Conference 2025*. HFES.

Volkova, S., Nguyen, D., Penafiel, L., Kao, H. T., Cohen, M., Engberson, G., Cassani, L., & Rebensky, S. (2025). VirTLab: Augmented intelligence for modeling and evaluating human-AI teaming through agent interactions. Manuscript submitted for publication.

Volkova, S., Rebensky, S., Cassani, L., McCormack, R., Fouse, A., Bruni, S., Gangberg, G., & Orvis, K. (2024). Compound AI ecosystem: Agents and tools to improve training and learning. Proceedings of the Interservice/Industry Training, Simulation and Education Conference.

W. N. Caballero and P. R. Jenkins, "On Large Language Models in National Security Applications," arXiv:2407.03453, 2024.

Walter, H., & Morrison, P. (2024). The evolution of autonomous tactical AI: Learning the lessons from Ukraine. Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC).

Xu, D., Chen, W., Peng, W., Zhang, C., Xu, T., Zhao, X., Wu, X., Zheng, Y., Wang, Y., & Chen, E. (2023). Large Language Models for Generative Information Extraction: A survey. arXiv preprint arXiv:2312.17617. https://arxiv.org/abs/2312.17617

Zaharia, M., Khattab, O., Chen, L., Davis, J. Q., Miller, H., Potts, C., Zou, J., Carbin, M., Frankle, J., Rao, N., & Ghodsi, A. (2024). The shift from models to compound AI systems. https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/

Zook A, Lee-Urban S, Riedl MO, Holden HK, Sottilare RA, Brawner KW. Automated scenario generation: toward tailored and optimized military training in virtual environments. In Proceedings of the international conference on the foundations of digital games 2012 May 29 (pp. 164-171).