xC2SAR: An Explainable and Auditable C2 Framework for Human–UAV Teaming in SAR Operations

Nikolaos Marios Militsis[1[0000−0001−9316−2348]], Achilleas Toumpas[1[0009−0008−6028−5622]], Ilias Koulalis[1[0000−0001−8268−2225]], Stéphane Zieba[2[0009−0004−7561−4119]], Edward-Benedict Brodie[2[0009−0007−0316−8027]], Konstantinos Ioannidis[1[0000−0001−6767−8762]], and Stefanos Vrochidis[1[0000−0002−2505−9178]]

[1] Centre for Research & Technology Hellas, Information Technologies Institute
{nikos.militsis,atoumpas,iliask,kioannid,stefanos}@iti.gr
[2] Thales cortAIx
{stephane.zieba,edward-benedict.brodie-of-brodie}@thalesgroup.com

Abstract. Unmanned Aerial Vehicles (UAVs) are increasingly deployed in Search and Rescue (SAR) operations because they can rapidly survey remote or difficult terrain and improve situational awareness in time-critical environments. However, the volume and heterogeneity of UAV data can also overload operators and hinder rapid decisions. Recent advances in Multimodal Large Language Models (MLLMs) enable joint reasoning over UAV imagery and mission text, but effective SAR support also requires structured domain knowledge and transparent, auditable outputs that operators can verify and trust. In this work, we propose xC2SAR, an explainable and auditable Command and Control (C2) framework for UAV-supported SAR operations that structures multimodal MLLM reasoning around the formal land-search strategies of the Australian National Search and Rescue Manual (NATSAR). The framework combines structured knowledge retrieval of SAR guidance, UAV imagery interpretation, and concept- and attribution-based explanations within a unified, auditable reasoning chain for human–UAV SAR teaming. We evaluate xC2SAR through a three-layer assessment combining scenario-based operational reasoning, quantitative explanation-quality metrics, and the Assessment List for Trustworthy AI (ALTAI). Code and data will be made available at `https://github.com/nikomili/xC2SAR`.

Keywords: Explainable AI · Decision Support · Search & Rescue · Unmanned Aerial Vehicles

1 Introduction

SAR operations are inherently complex, time-critical, and safety-sensitive, and are often conducted in uncertain and dynamically evolving environments. First responders and command personnel must make rapid decisions based on incomplete and changing information, while the consequences of error can be severe. In this context, UAVs are increasingly employed to extend the operational reach of response teams, provide aerial coverage of large or inaccessible areas, and enhance situational awareness. However, UAV platforms can also increase operator cognitive overload by generating large volumes of heterogeneous visual and contextual data that must be interpreted under time pressure [18]. Recent advances in MLLMs have created new opportunities for supporting SAR and C2 operations, as they enable joint reasoning over UAV imagery, textual scene descriptions, operational reports, and supporting documents. Such capabilities can support AI-assisted decision-making that not only identifies potentially relevant field cues but also contextualizes them for human operators.

Recent studies have explored MLLM-based support for UAV-enabled SAR through clue-analysis pipelines, cognitive guardrails, human-swarm teaming architectures, as well as broader system-level integrations of Large Language Models (LLMs) in UAV operations. In these frameworks, retrieved information typically consists of guidance, operational constraints, or domain knowledge used to support clue interpretation, task allocation, mission adaptation, or execution, while the generated outputs are commonly expressed as relevance assessments, rationales, task plans, concern reports, mission adaptations, or executable artifacts. However, existing approaches are not primarily designed to support formal SAR planning through retrieval of official search operational guidance, and they do not consistently provide a unified explainable and auditable chain linking clue assessment, retrieved knowledge, and downstream planning recommendations [2, 6, 18]. This limitation is especially important in missing-person incidents, where effective search planning depends on formal strategies for search-area determination and search-area coverage planning, as well as on transparent support that operators can inspect, verify, and trust. In safety-critical contexts, such transparency is also essential for reducing cognitive overload by transforming large volumes of multimodal information into prioritized and actionable decision support.

This work is further motivated by emerging requirements for transparency and human oversight in safety-critical AI systems. Perspectives from the European SEISMEC project [11] on human–AI symbiosis, together with industrial C2 assurance requirements explored with Thales cortAIx, informed an emphasis on auditability, operator oversight, and explainability as core design requirements. In this view, AI support for high-stakes SAR operations should provide verifiable and inspectable reasoning artefacts that support human judgment, enabling operators to challenge, verify, and integrate AI-generated evidence into operational decision-making.

Motivated by these limitations, we propose *xC2SAR*, an explainable and auditable C2 framework for UAV-supported land SAR decision support. The framework is designed around the formal land-search reasoning structure used in NATSAR, where theoretical, statistical, subjective/decision-point, and deductive strategies are combined to identify areas of higher probability. In the proposed workflow, UAV observations are not treated as autonomous triggers for replanning. Instead, they are treated as candidate evidence that may support, challenge, or refine the operator's interpretation of the incident when considered alongside retrieved SAR operational guidance, map context, missing-person information, and human judgement.

We demonstrate xC2SAR through a NATSAR-inspired missing-person scenario involving a 78-year-old person with dementia. The system uses EVisRAG [25] to retrieve and reason over textual and visual material from the NATSAR manual, including the scenario description, map evidence, Naismith-based movement assumptions, Lost Person Behaviour statistics, subjective boundary reasoning, and deductive clue guidance. We adapt the scenario by introducing a UAV-detected pair of gloves as candidate field evidence. This clue is assessed in the deductive stage, where concept-based explanations and attribution maps help expose whether the model has identified semantically meaningful visual evidence, such as glove-like shape, pair structure, colour, and clothing-related context. The resulting output is not an autonomous SAR decision, but an auditable review package for the Search and Rescue Mission Coordinator (SMC).

The main contributions of this work are as follows:

- We propose xC2SAR, an operational-guidance-grounded multimodal evidence-integration framework for UAV-supported SAR/C2 decision support that structures MLLM reasoning around formal land-search planning strategies aligned with NATSAR (theoretical, statistical, subjective and deductive).
- We introduce a unified, auditable reasoning workflow that organises multimodal inputs, retrieved SAR doctrinal guidance, and human judgement across theoretical, statistical, subjective/decision-point, and deductive strategies, producing traceable review packages for the SMC.
- We present a three-layer evaluation methodology tailored to explainable decision support, combining (i) scenario-based operational reasoning analysis, (ii) quantitative assessment of explanation quality (CLIPScore and concept overlap), and (iii) ALTAI-based trustworthiness self-assessment.

2 Background and Related Work

2.1 AI and UAVs for Search and Rescue Operations

Recent advances in LLMs and MLLMs are reshaping UAV-assisted SAR by improving perception, reasoning, mission planning, and human–machine interaction. By combining UAV mobility with language-model reasoning, these systems can support more adaptive and context-aware aerial operations, with the potential to improve task performance while reducing completion time and operator workload [10]. In high-stakes settings, prior work has also emphasised careful system design and calibrated autonomy to foster appropriate trust and human oversight [1].

At the drone-swarm level, traditional control and coordination architectures can increase operator workload due to large volumes of incoming information. Reviews of UAV swarm formation control therefore highlight the need for AI-assisted coordination in dynamic environments [4]. Human–swarm teaming frameworks incorporating LLMs have also been proposed for SAR, where LLM-based agents support intention understanding, hierarchical task decomposition, and mission planning, reducing operator workload and task completion time [18].

2.2 Multimodal Reasoning and Retrieval

MLLMs are relevant to UAV-assisted SAR because operators must jointly interpret aerial imagery, environmental context, textual instructions, and mission constraints. Recent UAV benchmarks evaluate these capabilities in increasingly realistic aerial settings. UAVBench evaluates agentic UAV systems [12]; SAREnv provides wilderness SAR scenarios, probability maps, and path-planning baselines [13]; MM-UAVBench evaluates perception, cognition, and planning in low-altitude UAV scenes [7]; AirCopBench focuses on multi-drone collaborative perception and decision-making under degraded conditions [27]; and DVGBench targets visual grounding in drone imagery [29].

These capabilities can be strengthened through retrieval-augmented generation (RAG), which allows MLLMs to incorporate external knowledge at inference time. In SAR settings, RAG can ground model outputs in operational guidance and contextual information, improving the interpretability and reliability of decision support [24].

2.3 Explainable AI in MLLMs

XAI methods have increasingly been applied to MLLMs to clarify how these models reason and what evidence supports their outputs. Because MLLMs process both visual and textual information, explanations may operate across token [16], embedding [21], layer [22], and neuron [20] levels. Existing approaches include causal, attention-based, concept-based, and surrogate-based explanations [8]. Attention and gradient-based methods can highlight image regions associated with model predictions [28], while concept-based methods use feature embeddings and techniques such as non-negative matrix factorisation (NMF) [23] and sparse autoencoders [5, 17] to identify human-interpretable concepts and their local activations.

3 Framework and Approach

We propose *xC2SAR*, an explainable and auditable C2 framework for UAV-supported land SAR decision support. The framework is motivated by the land-search planning structure described in the Australian National Search and Rescue Manual (NATSAR), where SAR personnel combine mission information, missing-person characteristics, map evidence, environmental constraints, and field observations to develop and reassess search strategies [3]. xC2SAR supports this process by retrieving relevant SAR operational guidance, organising heterogeneous evidence, and presenting reviewable reasoning traces to the Search and Rescue Mission Coordinator (SMC).

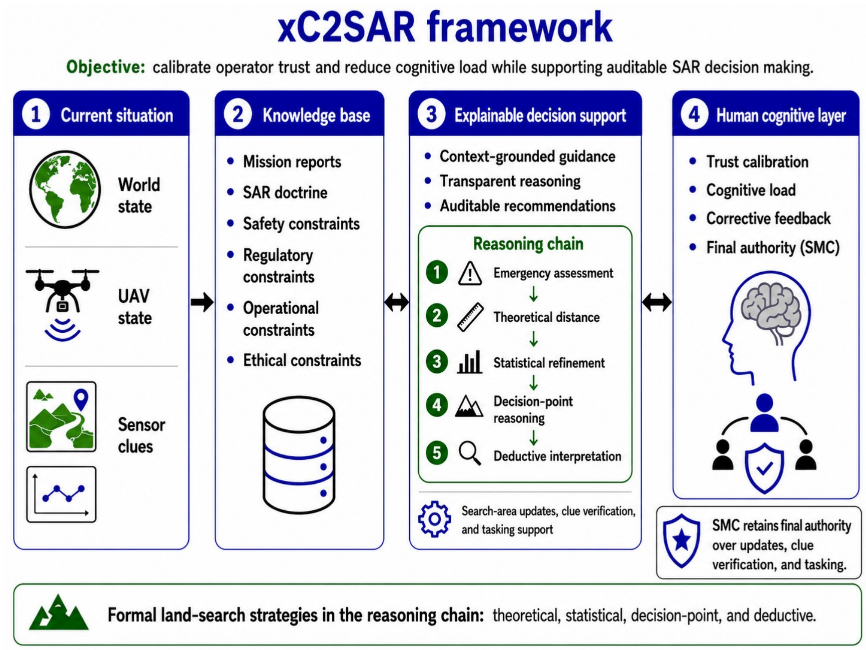


Fig. 1. Overview of xC2SAR as an operational-guidance-grounded multimodal evidence-integration framework for UAV-supported land SAR decision support. UAV observations are treated as evidence sources that may inform formal land-search reasoning but do not automatically update the search plan.

The framework remains advisory. It does not autonomously redefine the Last Known Position (LKP), Initial Planning Point (IPP), search area, asset allocation, or UAV tasking. Instead, it produces an auditable review package containing retrieved evidence, assumptions, uncertainty factors, explanations, and candidate planning implications. Any operational update requires confirmation by authorised SAR personnel.

Figure 1 summarises the proposed workflow. The framework is organised into three main components: inputs, a reasoning core, and an SMC review package. The inputs include: (i) mission context, such as LKP/IPP, elapsed time, terrain, weather, and available assets; (ii) missing-person information, such as age, condition, clothing, mobility, behavioural category, and vulnerability factors; and (iii) UAV-derived observations, such as imagery, detected objects, terrain cues, hazards, access routes, or candidate clues. At the core of the framework is an EVisRAG-based retrieval-augmented multimodal reasoning pipeline [25], which retrieves textual and visual evidence from SAR operational guidance and combines it with scene interpretation to support formal land-search reasoning.

The reasoning core is organised around five SAR reasoning components reflected in Figure 1: emergency assessment, theoretical, statistical, subjective/decision-point, and deductive reasoning. The **emergency-assessment component** summarises urgency-related evidence from the mission context and retrieved guidance for SMC review. The **theoretical strategy** retrieves scenario-specific information from NATSAR operational guidance and Naismith-based speed/time reference rules to estimate the theoretical search radius from the LKP or IPP. The **statistical strategy** retrieves Lost Person Behaviour evidence for the relevant missing-person category. The **subjective/decision-point strategy** uses map evidence, terrain features, and UAV observations to identify boundaries, barriers, attractors, routes, hazards, and likely decision points. The **deductive strategy** incorporates case-specific evidence, including sightings, reports, detected objects, footprints, clothing, belongings, heat signatures, or other possible clues.

In xC2SAR, UAV clue interpretation is part of the deductive strategy. A candidate clue is not assumed to belong to the missing person. Instead, the system assesses whether the object is visually consistent with the detected category, semantically consistent with the missing-person briefing, and operationally relevant given the wider context. Concept-based explanations and attribution maps expose the visual and semantic evidence supporting the clue assessment, while uncertainty notes and human-verification requirements preserve operator authority. The resulting outputs are structured for inclusion in the SMC review package rather than being used to automatically update the search plan.

4 Scenario Demonstration and Evaluation

4.1 Scenario-Based Proof of Concept

The proof of concept follows the structure of the land-search scenarios presented in NATSAR. The objective is to examine whether xC2SAR can follow a NATSAR-style land-search reasoning chain and provide adequate UAV-based clue interpretation and explainability in a controlled test scenario.

We use Scenario 1 from the NATSAR formal land-search strategy section. The original scenario describes a 78-year-old person with dementia missing from her home at 30 Brisbane St, Mt Ommaney. She was last seen at 0600, the report was received at 0900, and by 1200 she had been declared missing for three hours. The case includes important vulnerability factors: dementia, advanced age, poor walking ability, lack of footwear, no money or identification, light clothing, and cold wet weather [3]. NATSAR uses this scenario to illustrate how theoretical, statistical, subjective, and deductive strategies can be combined to identify a probable search area.

To exercise the multimodal deductive component, we introduce an adapted visual clue: a pair of grey gloves detected within the search area after a family member reports that the missing person often wore distinctive grey gloves in cold weather. The glove observation is treated as candidate deductive evidence, not as confirmed evidence, and does not automatically update the LKP, IPP, or priority search area.

The proof of concept proceeds in five stages. First, xC2SAR retrieves the scenario description, LKP map, and Search Urgency Assessment material to summarise urgency factors for the SMC, including age, cognitive condition, clothing, footwear, weather exposure, and elapsed time. This output is not an emergency-phase declaration; the SMC remains responsible for operational assessment and escalation.

Second, the theoretical strategy estimates possible distance travelled from the LKP. EVisRAG retrieves the NATSAR scenario material and Naismith-related speed guidance. Although a person could nominally walk about 5 km/h in suburban terrain, the retrieved evidence supports a reduced estimate of 1.5 km/h to account for age, dementia, poor mobility, and lack of footwear. Since three hours have elapsed, the theoretical distance is:

$$3 \text{ h} \times 1.5 \text{ km/h} = 4.5 \text{ km}.$$

xC2SAR presents this as a theoretical radius around the LKP, together with the retrieved evidence and assumptions.

Third, the statistical strategy retrieves Alzheimer's/dementia Lost Person Behaviour evidence, which indicates that 80% of missing persons in this category have been located within 3.2 km of the LKP. xC2SAR therefore identifies a 3.2 km statistical radius as a high-probability area for review, while noting that this is a prioritisation aid rather than an exclusion boundary.

Fourth, the subjective/decision-point strategy uses the scenario map and NATSAR guidance to identify features that may contain, redirect, attract, or constrain movement. In the scenario, relevant features include the Brisbane River, a creek, the Western Motorway, roads, tracks, watercourses, slopes, barriers, junctions, and access routes. xC2SAR marks these as planning-relevant features requiring operator interpretation.

Finally, the deductive strategy processes the UAV-detected gloves as candidate evidence. xC2SAR classifies the observation as *relevant but unconfirmed*: relevant because it is visually consistent with gloves, semantically consistent with the family report, and contextually relevant given the cold and wet conditions; unconfirmed because the image alone cannot establish ownership, recency, chain of custody, or displacement. The system recommends marking the geolocation as a priority investigation point if it lies within the statistical area or near a relevant decision point, pending ground verification.

Regarding transparency, concept-based explanations are used in this deductive clue-assessment stage. Their purpose is to help the operator inspect whether the model has identified meaningful visual and semantic evidence. Additionally, attribution maps complement this by showing the visual focus of the system when tasked with assessing clue relevance. Attribution maps are also generated for the retrieved material of the previous steps, to indicate model attention per document. These explanations are used to support review by the SMC and promote transparency and trust in the system.

This reasoning chain demonstrates the intended role of xC2SAR: retrieving and organising evidence across formal land-search strategies, adding explainable multimodal clue interpretation, and producing an auditable review package. At each stage, the SMC has final authority to accept, reject, revise, or disregard the recommendation.

Table 1. NATSAR-structured reasoning chain produced by xC2SAR during the scenario demonstration.

Stage	Retrieved or multimodal evidence	xC2SAR output for SMC review
Emergency assessment	Scenario description, vulnerability factors, weather, clothing, footwear, elapsed time	High-urgency evidence summary; SMC confirms operational urgency and response priority.
Theoretical strategy	Naismith-related speed guidance and scenario-specific mobility assumptions	4.5 km theoretical radius from LKP using 1.5 km/h over three hours; assumptions shown for review.
Statistical strategy	Lost Person Behaviour evidence for Alzheimer's/dementia cases	3.2 km statistical radius as a high-probability area; uncertainty retained because statistical guidance is not deterministic.
Subjective/decision-point strategy	Scenario map, terrain boundaries, roads, river, creek, motorway, tracks, slopes, decision points	Candidate containment boundaries, movement constraints, and decision points highlighted for operator interpretation.
Deductive strategy	UAV glove observation, missing-person briefing, environmental context, clue explanations	Glove classified as relevant but unconfirmed; geolocation recommended as a priority investigation point pending ground verification.

4.2 Explainable AI techniques and Models

Explainability in xC2SAR relies on two complementary approaches in the deductive clue stage: concept-based and attribution-based explanation. The concept-based component reveals *what* semantic patterns the model associates with a candidate clue, whereas the attribution-based component highlights *where* the supporting visual evidence is located in the image. Together, these techniques provide semantic and spatial evidence for clue interpretation, with their combination providing stronger evidence support compared to attribution-only explanations.

For concept-based explanation, we extract hidden-layer activations from the MLLM and apply NMF to identify interpretable latent concepts [23]. To generate the concept pool tailored to the clue type used in our scenario, we retrieved approximately 2,200 glove-related images from ImageNet [9] and used them to extract candidate glove concepts. We then manually selected 20 images, depicting gloves in forms relevant to missing-person incidents, such as gloves lying on the ground or appearing as potentially discarded personal items. One of these 20 images was chosen as the clue image used in the scenario, while the full set of 20 images was used for the quantitative evaluation of the explanation module. For each clue image, the framework reports the most activated concepts together with representative concept images and grounding words, enabling the operator to inspect whether the model's internal representation is consistent with the expected clue semantics.

The attribution-based component complements this semantic view with spatial evidence. To achieve this, we use two attribution methods: relative attention and gradient attention [28]. Both methods generate heatmaps over the input image, indicating the regions that most strongly influence the model. Relative attention highlights image regions that receive stronger model attention under the current missing-person clue prompt compared to a general image description prompt. Gradient attention reveals regions to which the prediction is most sensitive using gradient-weighted attention. These heatmaps help the operator verify whether the model focuses on the actual clue rather than irrelevant background elements.

Explainability is also applied to the retrieval-based stages of xC2SAR. In the theoretical, statistical, and emergency-assessment stages, we generate attention maps over retrieved evidence using the relative attention approach of [28], enabling the operator to inspect which regions of the retrieved pages, figures, or tables contribute most strongly to the system's reasoning. This allows the framework to provide not only a final recommendation, but also visual traces of how retrieved SAR guidance and scenario evidence support the overall reasoning chain.

Finally, our framework builds on the Qwen2.5-VL-7B-based instance used in EVisRAG [25], which we use throughout the reasoning and explainability pipeline. For the quantitative evaluation of concept-based explanations, we additionally compare against Qwen2-VL-7B [26] and LLaVA-1.5-7B [19].

4.3 Evaluation Scope and Metrics

The conducted evaluation is preliminary and scenario-based. It does not establish operational effectiveness in live SAR conditions, nor does it evaluate coverage planning, search-segment allocation, sweep-width estimation, or probability-of-detection optimisation. Instead, the evaluation focuses on whether xC2SAR can faithfully follow the formal NATSAR land-search reasoning structure while providing interpretable explanations and auditable review artefacts for the SMC.

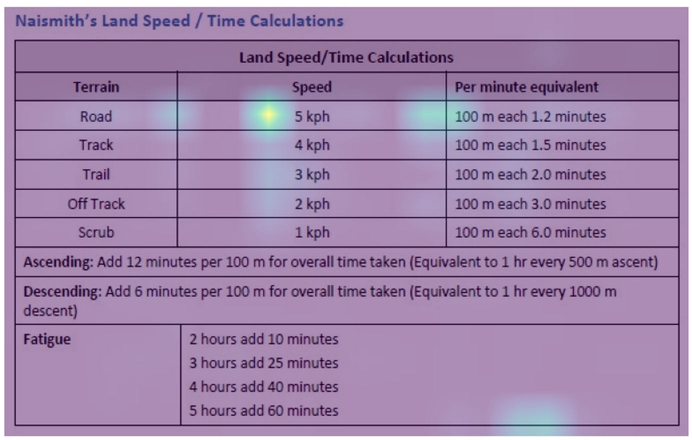
We evaluate the concept-based explanation component using two metrics. First, CLIPScore [14] assesses the correspondence between a UAV observation image and the grounded words associated with its top activated concepts. Following [23], this provides a label-free estimate of image–concept alignment. Second, concept overlap measures whether learned concepts capture distinct semantic information. Lower overlap indicates better disentanglement and supports the selection of the number of extracted concepts k in the NMF-based concept extraction process [23]. In addition, we assess broader trustworthiness through an ALTAI-based self-assessment [15].

4.4 Results

Quantitative explanation results Table 2 summarises the quantitative results for the concept-based explanation module. Lower concept-overlap values indicate better concept disentanglement, whereas higher CLIPScore indicates stronger

Table 2. Evaluation results for the three selected MLLMs. Concept overlap is reported for different numbers of concepts k; lower values indicate better disentanglement. CLIPScore evaluates alignment between clue images and grounded words from the top activated concepts; higher values indicate stronger image–concept alignment. Bold and underlined values indicate best and second-best performance for each metric, respectively.

Model	Concept overlap ↓			CLIPScore ↑
	$k = 5$	$k = 10$	$k = 20$	
LLaVA1.5	**0.09**	0.178	0.133	0.625
Qwen2	0.136	**0.08**	**0.04**	0.644
Qwen2.5	0.47	0.236	0.126	**0.654**


Naismith's Land Speed / Time Calculations

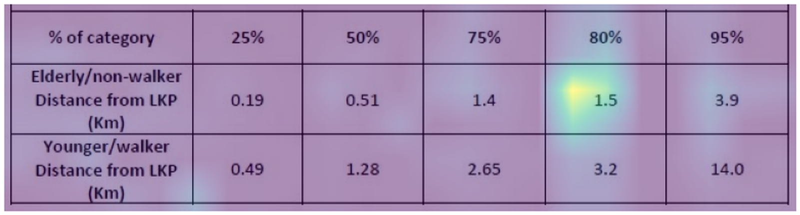
Land Speed/Time Calculations		
Terrain	Speed	Per minute equivalent
Road	5 kph	100 m each 1.2 minutes
Track	4 kph	100 m each 1.5 minutes
Trail	3 kph	100 m each 2.0 minutes
Off Track	2 kph	100 m each 3.0 minutes
Scrub	1 kph	100 m each 6.0 minutes
Ascending: Add 12 minutes per 100 m for overall time taken (Equivalent to 1 hr every 500 m ascent)		
Descending: Add 6 minutes per 100 m for overall time taken (Equivalent to 1 hr every 1000 m descent)		
Fatigue	2 hours add 10 minutes 3 hours add 25 minutes 4 hours add 40 minutes 5 hours add 60 minutes	

Fig. 2. Retrieved Naismith land speed/time evidence with attention overlay. The highlighted regions correspond to the movement-speed evidence used to support the theoretical search-area calculation.

alignment between UAV clue images and the grounded words of the most activated concepts.

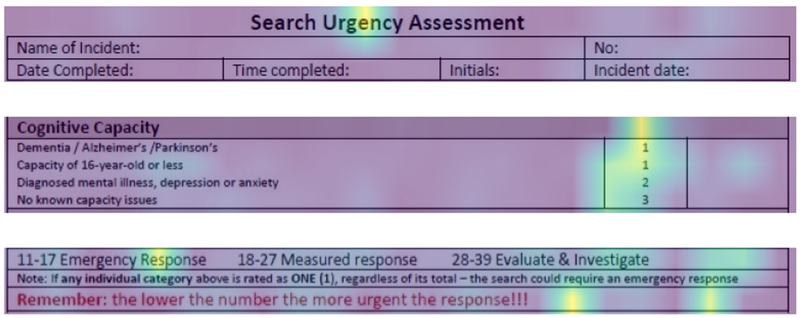
Qwen2.5 achieves the highest CLIPScore, suggesting the strongest image–concept alignment under this metric. Qwen2 achieves the lowest concept-overlap value, particularly for $k = 20$, suggesting more distinct concepts in this setting. The results indicate a trade-off between image–concept alignment and concept disentanglement. Since the differences in CLIPScore are small and the evaluation is limited to the explanation component, these results should be interpreted as preliminary evidence about explanation quality rather than operational SAR effectiveness.

Qualitative retrieval and explanation results Regarding explanation visualization, Figures 2 and 3 show retrieved NATSAR evidence used in the theoretical, statistical, and emergency-assessment stages. The attention maps for the retrieved



% of category	25%	50%	75%	80%	95%
Elderly/non-walker Distance from LKP (Km)	0.19	0.51	1.4	1.5	3.9
Younger/walker Distance from LKP (Km)	0.49	1.28	2.65	3.2	14.0

(a) Lost Person Behaviour evidence.



Search Urgency Assessment

Name of Incident:			No:
Date Completed:	Time completed:	Initials:	Incident date:

Cognitive Capacity		
Dementia / Alzheimer's /Parkinson's	1	
Capacity of 16-year-old or less	1	
Diagnosed mental illness, depression or anxiety	2	
No known capacity issues	3	

11-17 Emergency Response 18-27 Measured response 28-39 Evaluate & Investigate
Note: If any individual category above is rated as ONE (1), regardless of its total – the search could require an emergency response
Remember: the lower the number the more urgent the response!!!

(b) Search Urgency Assessment evidence.

Fig. 3. Retrieved evidence with attention overlays supporting statistical refinement and urgency assessment. In line with NATSAR, the 80% distance value is used because the remaining cases can skew higher-percentile estimates.

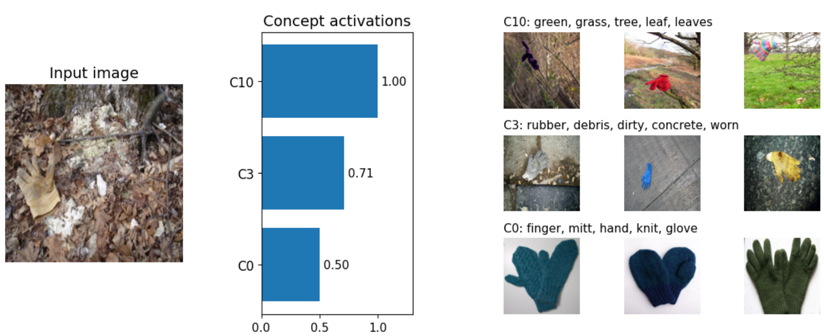
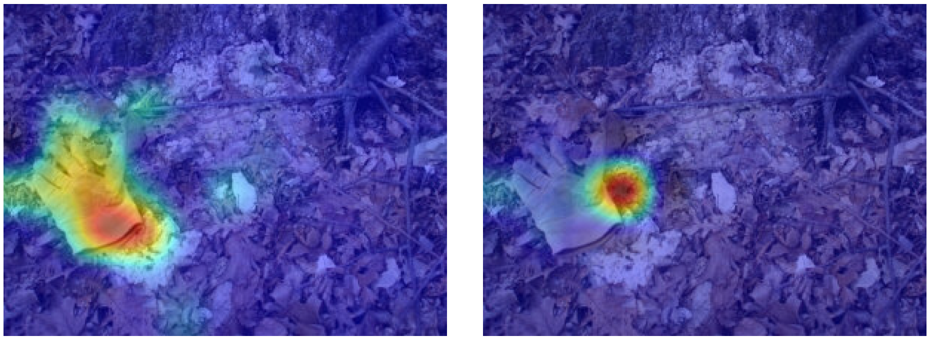


Fig. 4. Concept activation example for the deductive clue-assessment stage.

documents highlight scenario-relevant regions, suggesting that the model attends to evidence that is pertinent to the reasoning process.

Additionally, Figures 4 and 5 show explanation outputs for the deductive clue-assessment stage. The concept activation example shows three strongly activated concepts for the clue image, while the representative concept images and grounding words provide an interpretable view of the visual patterns captured by the model. In the heatmap examples, both methods localise attention to the glove on the ground, with the relative attention method producing clearer visual grounding in this example.



(a) Relative attention heatmap (b) Gradient attention heatmap

Fig. 5. Heatmap examples showing spatial regions emphasised during deductive clue interpretation.

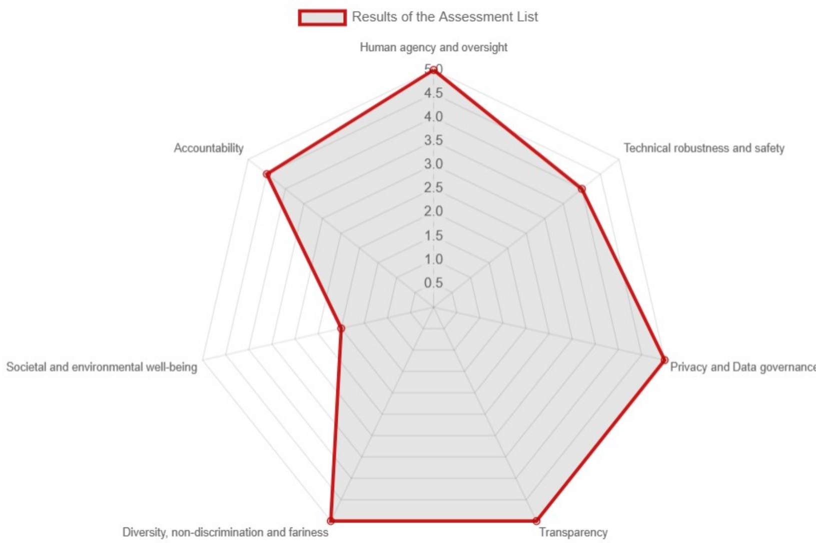


Fig. 6. ALTAI self-assessment results for xC2SAR across the seven Trustworthy AI requirements.

4.5 ALTAI Trustworthiness Self-Assessment

Because xC2SAR is intended for safety-critical SAR/C2 decision support, explanation quality alone is insufficient for assessing its suitability. We therefore use ALTAI as a structured self-assessment instrument, as a design checklist for identifying trustworthiness-related risks in our framework, as well as future governance requirements for operational readiness.

As shown in Figure 6, the assessment assigns high scores to human agency and oversight, privacy and data governance, transparency, diversity and fairness, and accountability. This aligns with the framework's advisory design, in which the SMC retains authority over LKP/IPP updates, search-area decisions, clue verification, and tasking. Technical robustness and safety receives a slightly lower score, highlighting the need for uncertainty communication, assumption tracing, hallucination detection, and validation under realistic SAR conditions. Societal and environmental well-being receives the lowest score, indicating that environmental impact, wider societal effects, and deployment-level consequences require further assessment. The reported ALTAI scores reflect a design-stage self-assessment conducted by the authors within the context of SEISMEC research project and provide preliminary indications of system trustworthiness.

Overall, the ALTAI assessment is deemed consistent with the intended role of xC2SAR as a human-supervised and auditable decision-support framework.

5 Conclusion

This paper presented *xC2SAR*, an explainable and auditable framework for UAV-supported SAR/C2 decision support. The framework combines multimodal reasoning, retrieval-augmented access to official SAR guidance, concept-based explanations, attribution maps, and operator-facing audit artefacts, while keeping the MLLM in an advisory role for review by the SMC and SAR personnel.

In an adapted NATSAR missing-person scenario involving a 78-year-old person with dementia, xC2SAR used EVisRAG to retrieve and reason over scenario text, map evidence, movement assumptions, Lost Person Behaviour statistics, terrain evidence, and deductive clue information. It organised this evidence into a NATSAR-style reasoning chain covering emergency assessment, distance estimation, statistical refinement, decision-point reasoning, and clue interpretation. A UAV-detected pair of gloves was treated as relevant but unconfirmed candidate evidence, with explanations helping the operator inspect the model's visual and semantic grounding.

The evaluation remains preliminary and scenario-based, and does not demonstrate operational effectiveness in live SAR missions or evaluate coverage planning and probability-of-detection optimisation. Future work will evaluate xC2SAR with SAR practitioners under realistic conditions, expand the benchmark with UAV clue imagery and map-based planning tasks, integrate geospatial tools for clue geolocation and operator-controlled search-area updates, and extend the framework toward sweep-width, coverage-factor, and probability-of-detection assessment.

Overall, xC2SAR contributes toward trustworthy, explainable, and auditable AI for UAV-supported SAR by linking multimodal clue interpretation, retrieved SAR guidance, formal land-search reasoning, and human-supervised decision authority.

Acknowledgements This project has received funding from the European Union's Horizon Europe research and innovation programme SEISMEC (grant agreement No 101135884).

References

1. Agrawal, A., Abraham, S.J., Burger, B., Christine, C., Fraser, L., Hoeksema, J.M., Hwang, S., Travnik, E., Kumar, S., Scheirer, W., et al.: The next generation of human-drone partnerships: Co-designing an emergency response system. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. pp. 1–13 (2020)
2. Alarcón Granadeno, P.A., Bernal Russell, A.M., Nelson, S., Hernandez, D., Petterson, M., Murphy, M., Scheirer, W.J., Cleland-Huang, J.: Cognition envelopes for bounded ai reasoning in autonomous uas operations. arXiv e-prints pp. arXiv–2510 (2025)
3. Australian Maritime Safety Authority: National Search and Rescue Manual. Australian Maritime Safety Authority, february 2026 edition edn. (Feb 2026), `https://www.amsa.gov.au/national-search-and-rescue-council/manuals-and-publications/national-search-and-rescue-manual`, available online only; latest edition. Accessed 28 April 2026
4. Bu, Y., Yan, Y., Yang, Y.: Advancement challenges in uav swarm formation control: A comprehensive review. Drones **8**(7), 320 (2024)
5. Bussmann, B., Nabeshima, N., Karvonen, A., Nanda, N.: Learning multi-level features with matryoshka sparse autoencoders. arXiv preprint arXiv:2503.17547 (2025)
6. Cleland-Huang, J., Granadeno, P.A.A., Bernal, A.M.R., Hernandez, D., Murphy, M., Petterson, M., Scheirer, W.: Cognitive guardrails for open-world decision making in autonomous drone swarms. arXiv preprint arXiv:2505.23576 (2025)
7. Dai, S., Ma, Z., Luo, Z., Yang, X., Huang, Y., Zhang, W., Chen, C., Guo, Z., Xu, W., Sun, Y., et al.: Mm-uavbench: How well do multimodal large language models see, think, and plan in low-altitude uav scenarios? arXiv preprint arXiv:2512.23219 (2025), `https://arxiv.org/abs/2512.23219`
8. Dang, Y., Huang, K., Huo, J., Yan, Y., Huang, S., Liu, D., Gao, M., Zhang, J., Qian, C., Wang, K., et al.: Explainable and interpretable multimodal large language models: A comprehensive survey. arXiv preprint arXiv:2412.02104 (2024)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848
10. Emami, Y., Zhou, H., Reddy, R., Arani, A.H., Wang, B., Li, K., Almeida, L., Han, Z.: Large language model-assisted uav operations and communications: A multifaceted survey and tutorial. arXiv preprint arXiv:2602.19534 (2026)
11. European Commission, CORDIS: Supporting european industry success maximization through empowerment centred development (SEISMEC) (2024). https://doi.org/10.3030/101135884, `https://cordis.europa.eu/project/id/101135884`, grant agreement ID: 101135884
12. Ferrag, M.A., Lakas, A., Debbah, M.: Uavbench: An open benchmark dataset for autonomous and agentic ai uav systems via llm-generated flight scenarios. arXiv preprint arXiv:2511.11252 (2025)
13. Grøntved, K.A.R., Jarabo-Peñas, A., Reid, S., Rolland, E.G.A., Watson, M., Richards, A., Bullock, S., Christensen, A.L.: Sarenv: An open-source dataset and benchmark tool for informed wilderness search and rescue using uavs. Drones (2025)
14. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: Proceedings of the 2021 conference on empirical methods in natural language processing. pp. 7514–7528 (2021)

15. High-Level Expert Group on Artificial Intelligence: The assessment list for trustworthy artificial intelligence (altai) for self-assessment. Tech. rep., European Commission (Jul 2020). https://doi.org/10.2759/002360, https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment
16. Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Wang, J., Lin, D., Zhang, W., Yu, N.: Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13418–13427 (2024)
17. Huben, R., Cunningham, H., Smith, L., Ewart, A., Sharkey, L.: Sparse autoencoders find highly interpretable features in language models. In: International Conference on Learning Representations. vol. 2024, pp. 7827–7845 (2024)
18. Ji, K., Hu, X., Zhang, X., Chen, J.: An llm-based framework for human-swarm teaming cognition in disaster search and rescue. arXiv preprint arXiv:2511.04042 (2025)
19. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
20. Meng, K., Bau, D., Andonian, A., Belinkov, Y.: Locating and editing factual associations in gpt. Advances in neural information processing systems **35**, 17359–17372 (2022)
21. Moayeri, M., Rezaei, K., Sanjabi, M., Feizi, S.: Text-to-concept (and back) via cross-model alignment. In: International Conference on Machine Learning. pp. 25037–25060. PMLR (2023)
22. Palit, V., Pandey, R., Arora, A., Liang, P.P.: Towards vision-language mechanistic interpretability: A causal tracing tool for blip. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2856–2861 (2023)
23. Parekh, J., Khayatan, P., Shukor, M., Newson, A., Cord, M.: A concept-based explainability framework for large multimodal models. Advances in Neural Information Processing Systems **37**, 135783–135818 (2024)
24. Sezgin, A., Boyacı, A.: Llm-powered uavs: a rag-based approach for safety-critical operations. In: International Conference on Intelligent and Fuzzy Systems. pp. 577–584. Springer (2025)
25. Sun, Y., Peng, C., Yan, Y., Yu, S., Liu, Z., Chen, C., Liu, Z., Sun, M.: Visrag 2.0: Evidence-guided multi-image reasoning in visual retrieval-augmented generation. arXiv preprint arXiv:2510.09733 (2025)
26. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution (2024), https://arxiv.org/abs/2409.12191
27. Zha, J., Fan, Y., Zhang, T., Chen, G., Chen, Y., Gao, C., Chen, X.: Aircopbench: A benchmark for multi-drone collaborative embodied perception and reasoning. arXiv preprint arXiv:2511.11025 (2025)
28. Zhang, J., Khayatkhoei, M., Chhikara, P., Ilievski, F.: MLLMs know where to look: Training-free perception of small visual details with multimodal LLMs. In: The Thirteenth International Conference on Learning Representations (2025), https://arxiv.org/abs/2502.17422
29. Zhou, Y., Chen, J., Huang, P., Ding, R., Zou, Z., Gao, P., Li, K., Yang, X., Jiang, X., Yang, H., Jonathan, L.: Dvgbench: Implicit-to-explicit visual grounding benchmark in uav imagery with large vision-language models. ISPRS Journal of Photogrammetry and Remote Sensing **232C**, 831–847 (2026)