

<https://doi.org/10.31891/csit-2026-3-22>

UDC 004.8:004.932

Yehor TERESHCHENKO

Bachelor of Engineering in Information
Technology, Staff Member
Metropolia University of Applied Sciences,
Helsinki, Finland

<https://orcid.org/0009-0009-1327-2834>

e-mail: Yehor.Tereshchenko@metropolia.fi

Mika HÄMÄLÄINEN

PhD in Language Technology, Development
Manager

Metropolia University of Applied Sciences,
Helsinki, Finland

<https://orcid.org/0000-0001-9315-1278>

e-mail: Mika.Hämäläinen@metropolia.fi

Received: 11/05/2026

Accepted: 28/07/2026

Published: 30/09/2026

© Copyright
2026 by the author(s)



This is an Open Access article distributed
under the terms of the [Creative Commons
CC-BY 4.0](https://creativecommons.org/licenses/by/4.0/)

HALLUCINATION-AWARE UAV DATASET BOOTSTRAPPING FOR ISR

This paper studies hallucination-aware annotation for UAV-style imagery under publication-safe conditions and treats the problem as one of measurement, reliability, and deployment rather than as a small prompt-engineering demonstration. The proposed framework combines ontology normalization, overlay masking, grounded summary generation, controlled prompting, and structured metadata capture so that labels, counts, and textual descriptions can be audited jointly. To address the gap between convenient multimodal demonstrations and publishable evidence, the study combines an archival audit of 7 historical military-style frames with a public proxy benchmark of 350 aerial images and frames drawn from AerialMPT, DLR-MVDA, and DLR-ACD. The executed public benchmark contains 274,581 ground-truth objects, moving the empirical basis of the work from a narrow pilot into a hundreds-scale evaluation. The local YOLO-World baseline completed all 350 public samples with mean latency 0.1807 s per image, but it under-counted dense scenes severely, with mean absolute count error of 143.889 on AerialMPT, 297.300 on full DLR-MVDA scenes, and 6848.242 on DLR-ACD. For a matched cloud-versus-local comparison, the paper introduces a dense-tile MVDA benchmark with 10 derived 1024x1024 crops and 315 labeled vehicles. A three-profile Gemini Robotics-ER ablation shows that prompt design materially changes outcome quality: the generic profile completed all tiles but over-counted heavily, the constrained ISR profile reduced error but dropped one request, and the best profile, coarse-to-fine, achieved MAE 5.300, MAPE 10.161%, and 10 / 10 completion on this slice. A supplementary video study on publicly available tactical footage then isolates text control: the freeform condition produced ontology expansion in 21 of 21 non-empty summaries, whereas the matched grounded condition reduced expansion to 0 of 20 while cutting latency from 491.070 s to 272.711 s. The archival audit further reveals unsupported unit, status, or intent language in 4 of 7 historical frames (57.1%) and four raw label variants (mil targets, military target, military targets, vehicle) for the same armored-platform concept. The contribution of the paper is therefore an experimentally grounded framework for hallucination-aware UAV dataset bootstrapping that reports semantic control, counting quality, prompt sensitivity, and execution robustness within one evaluation design.

Keywords: UAV ISR, annotation framework, grounded summaries, public aerial benchmark, open-vocabulary detection, secure local deployment.

1. Introduction

Rapid interpretation of UAV imagery has become a routine requirement in surveillance, infrastructure monitoring, emergency response, and military decision-support systems. In aerial-video workflows, converting raw footage into accurate, structured annotations and documented metadata can be more labor-intensive than acquisition, yet is essential for supervised model training, reproducible scientific use, and data reuse [1], [2]. Open-vocabulary detectors and multimodal annotators are attractive for this stage because they align visual evidence with text prompts instead of a fixed closed-set taxonomy [3], [4], [5].

For aerial imagery, however, annotation speed is not enough. Small object size, density, motion, oblique viewing angles, and background clutter are already difficult for conventional aerial benchmarks such as DOTA, UAVDT, and VisDrone [6], [7], [8]. When a vision-language or multimodal pipeline is allowed to produce free-form text in addition to boxes, another risk appears: the system may conflate scene evidence with overlaid text, language priors, or plausible but unsupported scene interpretation. Research on aerial image captioning and vision-language hallucination documents unsupported descriptions in aerial imagery and a tendency to over-rely on textual priors; a recent video-VLM preprint further reports systematic

hallucinations when on-screen text conflicts with the visual scene [9], [10], [11]. Hallucination studies in contemporary vision-language systems show that such unsupported content is systematic and measurable rather than anecdotal [12], [13], [14].

The problem considered in this article is therefore concrete: how can a prompt-driven Gemini Robotics-ER annotation framework be converted from a convenient labeling aid into a scientifically defensible pipeline for sanitized UAV-style data? The answer must address both output semantics and experimental accountability. A publishable pipeline should control the ontology, constrain the language emitted by the system, and report the reliability of both cloud and local execution paths.

The objective of the work is to design, implement, and validate such a pipeline with evidence that is large enough to be meaningful, yet publication-safe. The novelty is not a new detector architecture. The contribution lies in combining ontology normalization, grounded reporting, public proxy benchmarking, cloud-versus-local comparison, and explicit execution auditing within one empirical study. Instead of extrapolating from a narrow pilot, the paper establishes where the current approach already provides scientific and practical value and where further local adaptation is still required.

2. Related Work

YOLO-World [3] demonstrated that open-vocabulary detection can be implemented within a practical real-time detector family rather than being confined to slower vision-language stacks. For aerial imagery, the UAV-OVD line and the first dedicated review of open-vocabulary UAV detection show both the promise and the immaturity of the field: promptable detectors reduce ontology rigidity, yet small-object sensitivity, domain shift, and evaluation discipline remain unresolved [4], [5]. This trade-off is central to the present work because the proposed framework must remain useful on modest secure hardware while still accepting text-driven class constraints.

The dataset literature clarifies why aerial annotation pipelines require domain-specific caution. DOTA, UAVDT, and VisDrone became standard references because they expose the compound effects of density, motion, small target size, and viewpoint variation in overhead imagery [6], [7], [8]. AerialMPT extends this discussion into low-altitude pedestrian tracking [15], while DLR-ACD illustrates the extreme-density crowd regime in which even simple counting becomes difficult. Beyond the datasets executed in this study, newly released public DLR corpora such as EAGLE provide additional vehicle-centric evaluation material for future expansion [16]. These public benchmarks are valuable not only for model training, but also as publication-safe proxy workloads when private ISR data cannot be disclosed.

Hallucination analysis is now mature enough to inform evaluation design directly. HallusionBench measures entangled visual and linguistic hallucination [12], POPE-style protocols focus on unsupported object mentions [13], and VidHalluc extends the problem to temporal reasoning [14]. More recent mitigation studies such as visual evidence prompting and HalluSpace show that hallucination control is not only a benchmarking problem, but also an evidence-conditioning and representation problem [17], [18]. Their shared implication is methodological: unsupported language should be measured explicitly, not treated as an editorial issue. For annotation systems, this means that summaries and labels must be audited as outputs in their own right.

The final strand of related work concerns efficient secure deployment and label bootstrapping. RT-DETR shows that modern detectors can remain competitive while improving structural efficiency [19]. Unbiased Teacher, Soft Teacher, and Active Teacher demonstrate that pseudo-label quality, consistency control, and uncertainty-aware sample selection materially influence student detector performance [20], [21], [22]. This supports a two-phase architecture in which cloud models are used only for publication-safe bootstrap labeling, while later offline deployment relies on smaller local detectors and deterministic reporting. The present paper stops short of training the planned local student, but it provides the empirical baseline that such future distillation work must exceed.

3. System Architecture

The study uses a two-branch experimental pipeline in which sanitized media are ingested with Unicode-safe handling, evaluated under predefined prompt profiles, exported in YOLO [23] and COCO [24] formats, and stored together with structured prediction metadata. This design separates exploratory use from the controlled experimental pathway adopted for evaluation.

The use of structured metadata for provenance and auditing follows the same general traceability principle as prior work on an LLM reporting pipeline that versioned its inputs and retained source, date, category, and version-chain metadata for generated summaries [25].

The ontology is constrained to eight classes: `personnel`, `civil_vehicle`, `truck`, `armored_vehicle`, `tank`, `artillery_or_launcher`, `uav`, and `fortification_or_emplacement`. Earlier baseline outputs exhibited uncontrolled variation such as `vehicle`, `military target`, `military targets`, and `mil targets` for the same armored-platform concept. Ontology normalization is a deterministic controlled-vocabulary step that lowercases labels, matches them against a manually defined synonym table, replaces recognized variants with canonical classes, and rejects unknown labels before export, counting, or grounded text generation.

Text generation is further constrained through grounding. Free-form descriptions are retained only for diagnostic analysis, whereas the scientific path uses detection-conditioned text derived exclusively from accepted

boxes. Overlay masking can additionally suppress telemetry bars or corner overlays before cloud inference. In combination, these controls transform a convenient labeling interface into a measurement-oriented annotation framework.

Figure 1 presents the resulting architecture. Public or publication-safe media enter the experimental pipeline, pass through ontology and grounding controls, and are then processed by either the cloud annotator or the offline local baseline. The central methodological choice is that scientific claims are tied to stored predictions and quantitative summaries rather than to retrospective narrative interpretation.

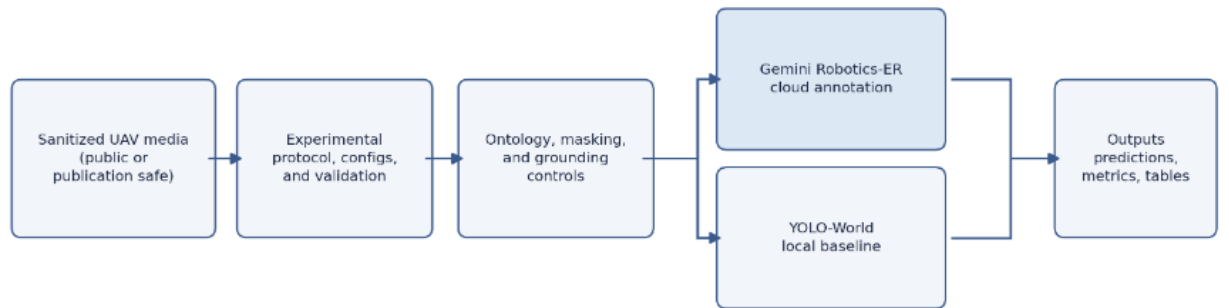


Figure 1. Experimental design of the evaluated cloud-and-local annotation framework.

4. Dataset and Annotation Protocol

The data used in this study are divided into four groups. The first group contains seven military-style frames processed in six earlier test runs of the annotation pipeline; one video run contributed two sampled frames. For each frame, the stored output includes detection boxes, raw labels, and a free-form summary. These outputs are re-examined to identify inconsistent labels and statements unsupported by the detections, with representative examples shown in Figures 7 and 8. They are not used for count benchmarking. The second group is a public proxy benchmark consisting of AerialMPT, DLR-MVDA, and DLR-ACD. These datasets were selected because they are openly obtainable, clearly licensed for research use, and rich in the small-object regimes that make UAV annotation difficult. The third group is a qualitative military pilot containing sanitized tank and convoy scenes, which are retained to verify that the ontology-constrained pipeline still behaves sensibly on ISR-like imagery. The fourth group is a supplementary local video case study based on publicly available tactical footage (clips aggregated from multiple public releases attributed to Ukrainian defence and security institutions and from community-archived social-media sources), used only for qualitative analysis and appendix material rather than for the main public benchmark claims.

The public proxy benchmark contains 350 images or video frames that were processed by the local baseline and included in the evaluation, with 274,581 ground-truth objects in total. AerialMPT contributes 307 frames from public aerial pedestrian-tracking sequences. DLR-MVDA contributes 10 annotated overhead traffic scenes with 3,505 labeled vehicles. DLR-ACD contributes 33 extreme-density crowd images with 226,336 point annotations used here for dense person counting. License verification and acquisition relied on the official DLR dataset pages for AerialMPT, DLR-MVDA, and DLR-ACD [26], [27], [28]. For the cloud comparison, the paper additionally derives ten 1024x1024 DLR-MVDA tiles that concentrate the dense vehicle regions of the full scenes. This derived tile set contains 315 labeled vehicles and is small enough to be processed within the practical limits of the cloud API.

This design serves three purposes. First, it allows the paper to report executed results on hundreds of public aerial samples instead of extrapolating from a two-scene pilot. Second, it preserves a clean ethical and reproducibility boundary: only public or publication safe imagery enters the cloud loop for the core benchmark claims. Third, it keeps a separate place for local-only qualitative material such as the open-source drone montage described in Appendix A, which can enrich the appendix without being conflated with the public benchmark. A natural next benchmark extension is EAGLE [16], which offers a larger vehicle-centric public workload under the same DLR distribution model.

Table 1.

Dataset sources, evaluated image and frame counts, and licensing status

Source	Samples	GT objects	License / access	Role in manuscript
Archival audit subset	7	23	Prior experimental outputs	Measures unsupported summary content and vocabulary drift in earlier unconstrained outputs
AerialMPT	307	44,740	CC BY-NC-ND 3.0	Primary public proxy benchmark for low-altitude personnel counting in aerial sequences
DLR-MVDA	10	3,505	CC BY-NC-ND 3.0	Primary public proxy benchmark for vehicle-heavy overhead scenes and cloud-vs-local comparison
DLR-MVDA dense tiles	10	315	Derived from DLR-MVDA	Curated 1024x1024 crops used for the executed Gemini tile benchmark
DLR-ACD	33	226,336	CC BY-NC-ND 3.0	Extreme-density crowd stress test for public small-object counting
Open-source drone montage (appendix case study)	24	N/A	Publicly available footage processed locally	Supplementary qualitative appendix case; excluded from core public benchmark claims

5. Experimental Design

The evaluation has five tracks. First, a retrospective audit examines seven stored frames from six earlier unconstrained pipeline runs. For each frame, the raw detection labels are mapped to the eight canonical classes, and the associated free-form summary is checked for unit, status, or intent statements that are not supported by the stored detections. This diagnostic audit measures vocabulary drift and narrative overreach; it is not used to estimate detection accuracy. Second, the local secure baseline runs YOLO-World across the full public benchmark of 350 images and frames using dataset-specific prompts: person and pedestrian for AerialMPT and DLR-ACD, and car, vehicle, van, bus, and truck for DLR-MVDA. Third, the cloud comparison runs Gemini Robotics-ER on the derived MVDA tile subset. Fourth, a matched three-profile prompt ablation compares generic, constrained ISR, and coarse-to-fine prompting on exactly the same ten dense public tiles. Fifth, a summary-control experiment on the supplementary open-source drone montage compares free-form scene descriptions against grounded summaries derived from the accepted detections.

The primary quantitative target in the public benchmark is count accuracy rather than mean average precision (mAP), because the datasets span boxes, tracking annotations, and point-level crowd marks. In the results tables, N denotes the number of evaluated images, frames, or tiles; GT obj. and Pred. obj. denote the total numbers of ground-truth and predicted objects; MAE denotes mean absolute error; MAPE denotes mean absolute percentage error; and Latency s denotes the mean processing time per sample in seconds. These measures provide a consistent comparison across full scenes and dense crops without hiding failure cases inside a single aggregate detection score.

For the retrospective audit, a phrase is counted as unsupported when it names a unit, operational status, or intent that is absent from the stored detection labels. The check uses a predefined keyword list and therefore identifies explicit lexical inconsistencies rather than judging every possible interpretation of the image. Although simpler than HallusionBench, POPE, or VidHalluc [12], [13], [14], it is adequate for identifying the specific vocabulary drift and narrative additions present in these earlier outputs. The grounded-summary mode is then evaluated as a control mechanism: if a class is absent from the detections, it should not appear in the final sentence.

The choice of baselines is intentionally conservative. YOLO-World represents an immediately deployable local open-vocabulary fallback on a laptop-class GPU [3]. Gemini Robotics-ER represents the prompt-driven cloud annotator under study. The comparison is therefore not framed as cloud versus the best possible local detector, but as cloud bootstrap versus an available secure fallback. The broader research roadmap still includes student training with RT-DETR or a comparable compact detector and a pseudo-label refinement loop informed by semi-supervised teacher-student methods [19], [20], [21], [22]. The present results are meant to define that target rather than to conclude the deployment story.

6. Results and Discussion

The most important empirical change relative to the earlier pilot is scale. The revised study now evaluates 350 public aerial samples with 274,581 ground-truth objects, rather than relying on a pair of illustrative frames. This broader evidence shows that the immediate local fallback is operationally attractive but quantitatively weak in dense small-object regimes. YOLO-World completed the full benchmark with mean latency 0.1807 s per sample, yet its count error remained high across all datasets: MAE 143.889 on AerialMPT, 297.300 on full DLR-MVDA scenes, and 6848.242 on DLR-ACD. These dataset-level results are summarized in Table 2. The pattern is consistent: the local open-vocabulary baseline preserves locality and speed, but it underestimates dense targets by a large margin.

The DLR-MVDA dense-tile comparison provides the clearest cloud-versus-local evidence under matched public conditions. On all ten tiles, the local baseline obtained MAE 14.300 and MAPE 48.939%. The best Gemini Robotics-ER configuration, coarse-to-fine, produced substantially lower tile-level error: MAE 5.300 and MAPE 10.161% on 10 completed tiles containing 315 labeled vehicles. In this revised experiment suite, the best cloud profile completed 10 of 10 dense-tile requests, so the main trade-off shifted from completion failure toward latency and prompt sensitivity.

The prompt ablation itself is scientifically important because it shows that cloud annotation quality is not a fixed property of the backend. The generic profile completed all ten tiles but over-counted severely, producing 650 predicted vehicles for 315 ground-truth vehicles and MAE 36.700. The constrained ISR profile reduced the error to MAE 8.111, but dropped one request in the tested environment. The coarse-to-fine profile combined full completion with the best count quality, achieving MAE 5.300 and MAPE 10.161%. This establishes prompt design as an experimental variable that must be reported explicitly in peer-reviewed annotation studies.

The retrospective review captures a failure mode that pure count metrics miss. Unsupported unit, status, or intent language appeared in 4 of 7 historical frames (57.1%), and four uncontrolled raw labels (mil targets, military target, military targets, vehicle) described the same armored-platform concept before ontology normalization. Representative predictions from these earlier outputs are shown in Figures 7 and 8. This confirms that label instability and narrative overreach were recurring properties of the examined unconstrained outputs.

The qualitative military pilot clarifies how the ontology-constrained cloud path behaves on sanitized ISR-like imagery. On the snow-forest scene, the constrained cloud path returned 1 tank. On the convoy scene, it returned 11 armored_vehicle, 1 truck. These cases are not part of the main quantitative benchmark, but they show that the same ontology and grounded reporting controls remain compatible with sanitized military-style scenes. The local pilot

remained faster (5.5988 s on average) than the cloud qualitative condition (9.4753 s), but it was visibly less semantically specific. Representative cloud and local predictions, together with the overlay-control cases, are shown in Figure 6 in Appendix B.

The supplementary open-source drone montage case study extends this qualitative analysis from isolated frames to a short edited sequence. At matched sampling of 24 frames, the freeform Gemini condition completed without frame errors and returned 52 ontology detections in 491.070 s, whereas YOLO-World returned 83 detections in 88.461 s. The most important observation is not the raw count difference alone, but the text behavior: all 21 non-empty Gemini summaries added actions, scene context, or object attributes beyond the closed ontology labels. Representative frame-level predictions from this montage are shown in Figures 3--5 in Appendix A.

A matched grounded-summary rerun on the same montage shows that this behavior can be controlled rather than merely observed. The grounded run completed all 24 frames, produced 49 detections, and reduced summary expansion to 0 of 20 non-empty summaries, while lowering end-to-end latency by 218.359 s (44.466%). The grounded mode is less descriptive, but it is analytically safer and cheaper to execute. This makes the freeform mode useful for appendix-style qualitative interpretation, while grounded mode is the defensible default for scientific reporting.

Taken together, the executed results support a stronger and more defensible claim than the earlier pilot allowed. Prompt-driven cloud annotation can improve count quality on vehicle-heavy overhead crops, but the result depends materially on prompt structure, and descriptive summaries require explicit grounding control. Immediate local open-vocabulary deployment remains robust and fast, but it is not yet adequate for dense small-object annotation without further adaptation. A scientifically useful UAV annotation framework therefore needs both prompt ablation reporting and grounded text generation, alongside a staged path from public cloud-assisted bootstrap labeling to later local secure deployment.

Table 2.

Public count benchmark and tile-level cloud comparison

Slice	Backend	N	GT obj.	Pred. obj.	MAE	MAPE %	Latency s
AerialMPT	YOLO-World	307	44,740	566	143.889	98.455	0.1467
MVDA full scenes	YOLO-World	10	3,505	532	297.300	82.470	0.3724
DLR-ACD	YOLO-World	33	226,336	344	6848.242	99.543	0.4391
DLR-MVDA dense tiles	YOLO-World	10	315	186	14.300	48.939	0.2640
MVDA dense tiles	Gemini ER (coarse to fine)	10	315	280	5.300	10.161	78.2101

Table 3.

Execution speed and completion audit

Backend	Slice	Completion	Latency s	Comment
YOLO-World	350 public samples	350 / 350	0.1807	Stable sub-second inference across public proxy workloads, but severe undercounting of dense small objects
YOLO-World	10 MVDA dense tiles	10 / 10	0.2640	Fast and reliable on cropped vehicle scenes, but still undercounts relative to ground truth
Gemini Robotics-ER	10 MVDA dense tiles (coarse to fine)	10 / 10	78.2101	Best prompt profile in the ablation; accuracy improved substantially over generic prompting at the cost of higher latency
Gemini Robotics-ER	2 sanitized military pilot scenes	2 / 2	9.4753	Constrained military prompts remained useful for qualitative tank and convoy scenes
YOLO-World	Open-source montage at 0.2 fps	24 / 24	3.6859	Local baseline on the supplementary video case study with full completion and grounded labels only
Gemini Robotics-ER	Open-source montage at 0.2 fps	24 / 24	20.4612	Freeform scene descriptions available for all sampled frames, but at much higher latency than the local baseline
Gemini Robotics-ER	Open-source montage, grounded summaries	24 / 24	11.3629	Grounded wording removed scene-level expansion and cut video-level latency substantially compared with freeform summaries

Table 4.

Archival hallucination and robustness audit			
Condition	Observed issue	Count/value	Interpretation
Archival summaries	Frames with unsupported unit/status/intent phrases	4 / 7	Historical free-form summaries often incorporated overlay text or unsupported operational context
Archival labels	Raw label variants for the same armored-platform concept	4	vehicle, military target, military targets, and mil targets were normalized to armored vehicle
Public proxy benchmark scale	Executed local benchmark workload	350 samples / 274,581 GT objects	The revised study is based on hundreds of public aerial samples instead of a two-scene pilot only
MVDA dense-tile cloud benchmark	Successful Gemini requests	10 / 10	Cloud completion, not only count accuracy, must be reported as a systems result
Grounded military pilot summaries	Unsupported cue hits	0 in executed pilot scenes	Detection-conditioned wording remained conservative on the qualitative tank and convoy examples
Montage freeform summaries	Frames where Gemini expanded beyond the closed detection ontology	21 / 21	All non-empty supplementary video summaries added scene-level actions or context beyond the detector labels
Montage grounded summaries	Frames where grounded summaries expanded beyond the closed detection ontology	0 / 20	Detection-conditioned wording removed unsupported scene-level elaboration in the matched grounded run

Table 5.

Gemini prompt-profile ablation on MVDA dense tiles							
Profile	Completion	GT obj.	Pred. obj.	MAE	MAPE %	Latency s	Interpretation
generic	10 / 10	315	650	36.700	51.757	49.2034	Open vocabulary prompt completed all tiles but over-counted vehicle clusters heavily.
constrained ISR	9 / 10	225	276	8.111	24.240	57.1221	Ontology control reduced overcounting, but one dense tile still failed in the tested environment.
coarse to fine	10 / 10	315	280	5.300	10.161	78.2101	Two-stage prompting delivered the best count accuracy and full completion on the public tiles.

Table 6.

Open-source montage summary-mode control experiment						
Summary mode	Frames	Detections	Expansion frames	Latency s	Seconds/frame	Interpretation
Freeform	24	52	21 / 21	491.070	20.4612	Richer narrative descriptions, but every non-empty summary expanded beyond the closed ontology.
Grounded	24	49	0 / 20	272.711	11.3629	Grounded generation preserved box-conditioned wording and removed scene-level elaboration.

7. Security and Deployment Discussion

The deployment implication of the benchmark is a two-stage research and deployment strategy rather than a single winning model. Public or publication-safe media can be routed through the cloud annotator to accelerate bootstrap labeling on the slices where cloud completion and accuracy are acceptable. The corrected outputs from that stage should then be used to train or fine-tune smaller local detectors for later offline operation. In other words, the cloud path is best treated as an accelerator for dataset creation, not as the final secure endpoint.

At the generative reporting layer, retaining human review is also consistent with prior comparative safety testing of LLMs, which found uneven safeguards across model families and emphasized robust oversight in high-stakes use [29].

This architecture is important for both security and scientific hygiene. Because the paper's quantitative claims are based on public proxy datasets and sanitized military examples, the evidence remains reproducible and publication-safe. Private or operational media do not need to enter the cloud loop to validate the central research claim. At the same time, the measured weakness of the immediate local fallback shows why an eventual local student model remains necessary for security-sensitive deployment.

The benchmark also clarifies where future local alternatives should be targeted. The undercounting observed on AerialMPT and DLR-ACD means that local secure deployment cannot rely on a generic off-the-shelf open-vocabulary detector alone. Future work should therefore combine corrected pseudo-labels from the cloud-assisted bootstrap stage with compact detectors such as RT-DETR-class students, domain-tuned counting heads, and deterministic templated reports, following the teacher-student and sample-selection principles highlighted in recent semi-supervised detection research [19], [20], [21], [22]. Figure 2 summarizes this security split between public cloud-assisted research and later local-only operational use.

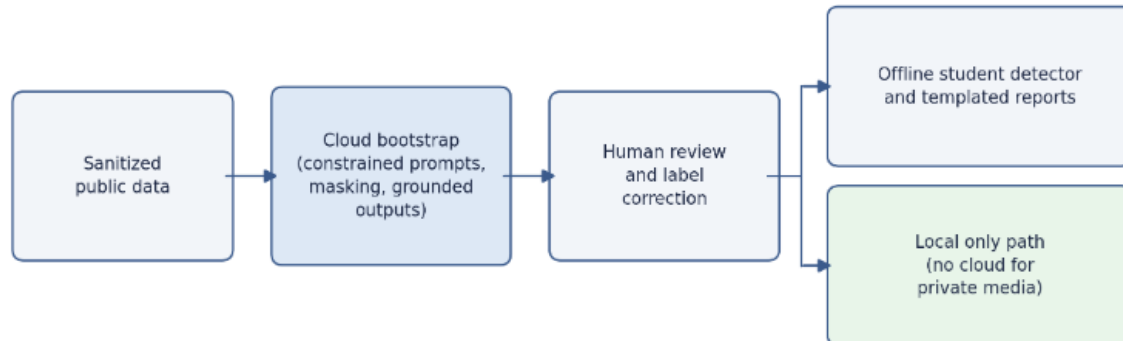


Figure 2. Proposed two-phase security architecture for later large-scale deployment

8. Conclusions

The study shows that hallucination-aware UAV annotation becomes scientifically credible only when it is evaluated as a measured system rather than as an informal prompt demonstration. After introducing ontology control, grounded reporting, and reproducible experimental logging, the empirical basis of the paper increased from a tiny pilot to 350 evaluated public aerial images and frames containing 274,581 ground-truth objects, plus a targeted archival audit of historical military-style outputs.

Four findings are most important. First, earlier unconstrained outputs genuinely suffered from both unsupported narrative content and unstable label vocabulary. Second, the immediately deployable local open-vocabulary baseline is fast and reliable but substantially undercounts dense small objects, so locality alone is not enough. Third, the cloud annotator can produce much better vehicle counts on dense public tiles, but prompt structure is a first-class determinant of quality: coarse-to-fine outperformed both generic and constrained single-stage prompting in the executed ablation. Fourth, freeform scene descriptions are useful for qualitative interpretation, yet the matched summary-control experiment on the open-source montage showed that grounded summaries materially reduce narrative expansion and latency. The practical conclusion is therefore not that one backend has won, but that publication-safe cloud bootstrap labeling, grounded reporting, and later local adaptation are complementary stages of one evaluation and deployment strategy.

9. Appendix A. Supplementary Open-Source Drone Montage Case Study

A supplementary local case study was executed on publicly available tactical footage (116.6 s). The footage aggregates clips from multiple public releases attributed to Ukrainian defence and security institutions and from community-archived social-media sources. It was sampled at 0.2 fps into 24 analyzed frames. This material is not part of the public benchmark and is reported separately as qualitative evidence only. Gemini Robotics-ER completed all sampled frames without frame-level API failures, producing 52 ontology detections and a freeform scene description for every frame. YOLO-World completed the same schedule locally with 83 detections.

The timing contrast in this local case study follows the same pattern as the public experiments. The freeform Gemini run required 491.070 s in total (20.4612 s per frame), the grounded Gemini rerun required 272.711 s (11.3629 s per frame), and YOLO-World required 88.461 s (3.6859 s per frame). Freeform Gemini summaries were more expressive, but all 21 non-empty summaries expanded beyond the closed ontology labels by adding contextual actions, scene attributes, or inferred object functions, whereas the grounded rerun reduced this value to 0 of 20.

Appendix A presents 18 carefully filtered representative frames spanning close-range personnel movement, hillside blast cues, nighttime breaching, dense personnel formation, UAV-related scenes, artillery or heavy-weapon cues, urban movement, armored-platform context, helicopter insertion, and thermal imagery. Frames that were visually weak, nearly empty, or redundant were intentionally excluded. The figure panels preserve the Gemini-labeled boxes and embed compact scene descriptions directly into the cards so that each example can be reviewed without leaving the appendix.

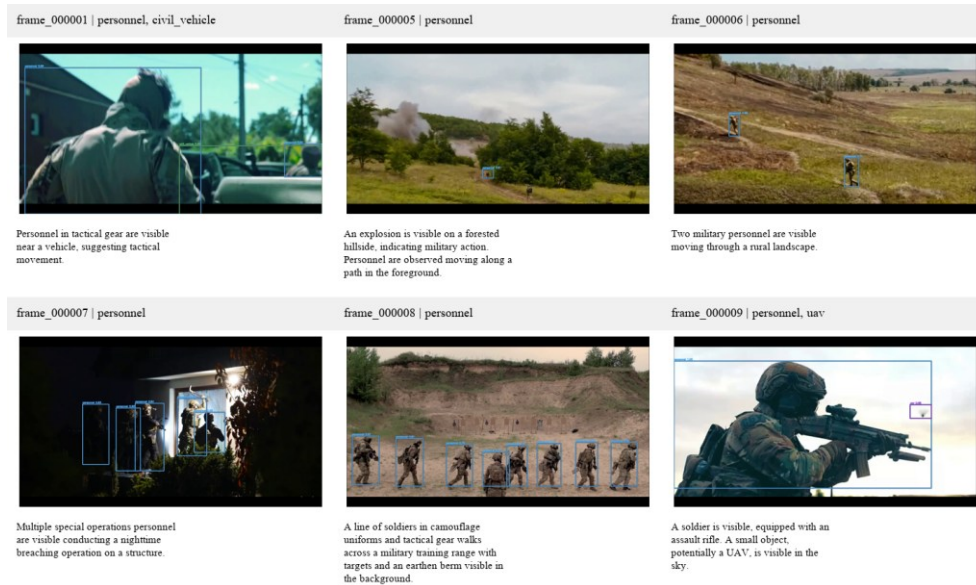


Figure 3. First six curated Gemini-labeled representative frames from the supplementary open-source drone montage.

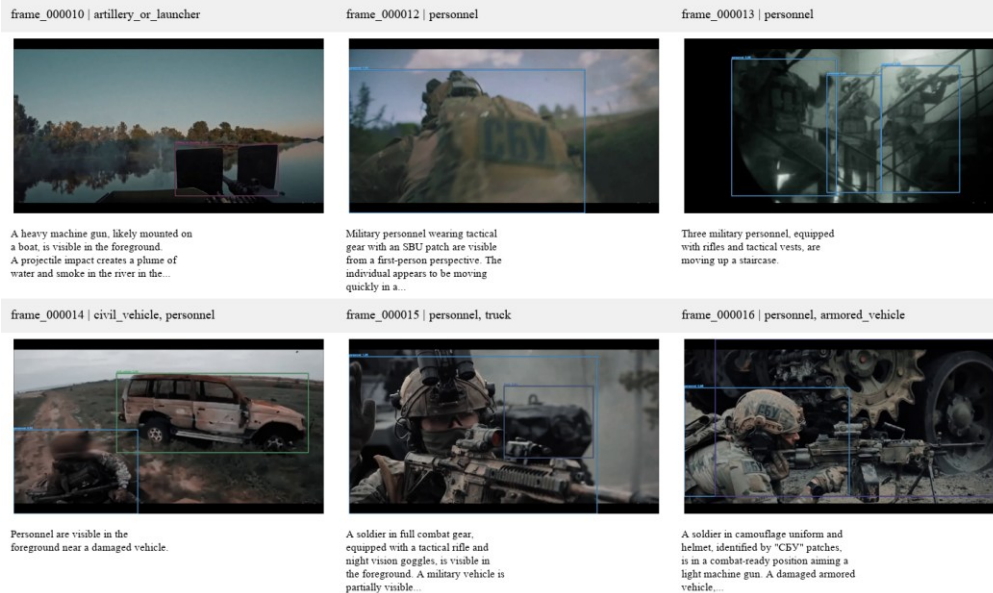


Figure 4. Second set of six curated Gemini-labeled representative frames from the supplementary open-source drone montage.

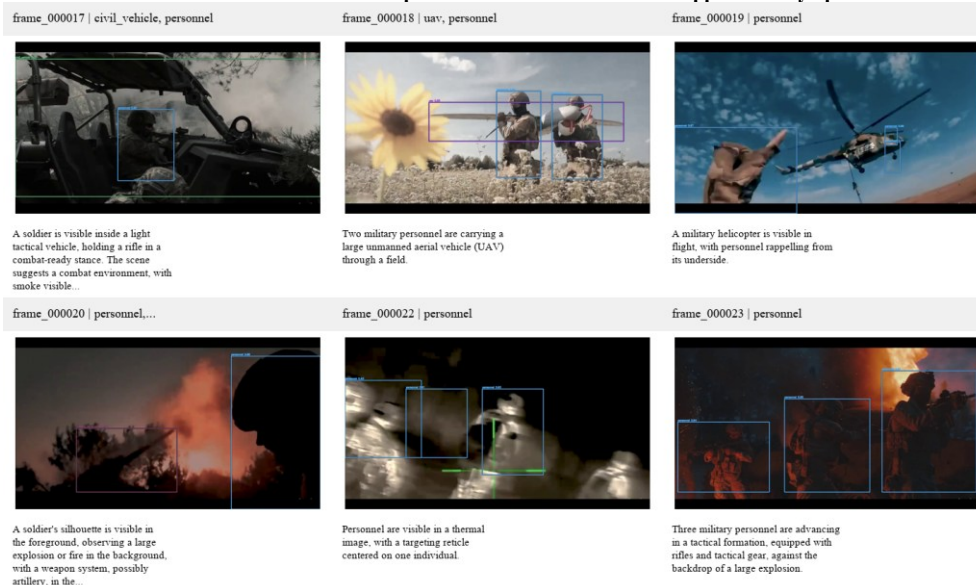


Figure 5. Third set of six curated Gemini-labeled representative frames from the supplementary open-source drone montage.

10. Appendix B. Supplementary Qualitative Comparison Cases

Appendix B assembles supplementary qualitative comparison cases in three groups: sanitized pilot scenes, overlay-control scenes, and archival baseline outputs. These panels are included to support visual interpretation of the main findings rather than to introduce additional headline quantitative claims.

Panels P1-P4 compare the cloud and local baselines on sanitized tank and convoy scenes. Panels C1-C2 contrast overlay-visible and overlay-masked conditions on the same FPV road scene in order to illustrate the effect of text overlays on scene description.

Panels A1-A4 and A5-A8 present archival baseline outputs drawn from earlier unconstrained conditions. Together these curated panels illustrate vocabulary drift, overlay-driven narrative contamination, markdown-style textual output, and cross-case variation among labels such as vehicle, military target, military targets, mil targets, and tank before ontology normalization.

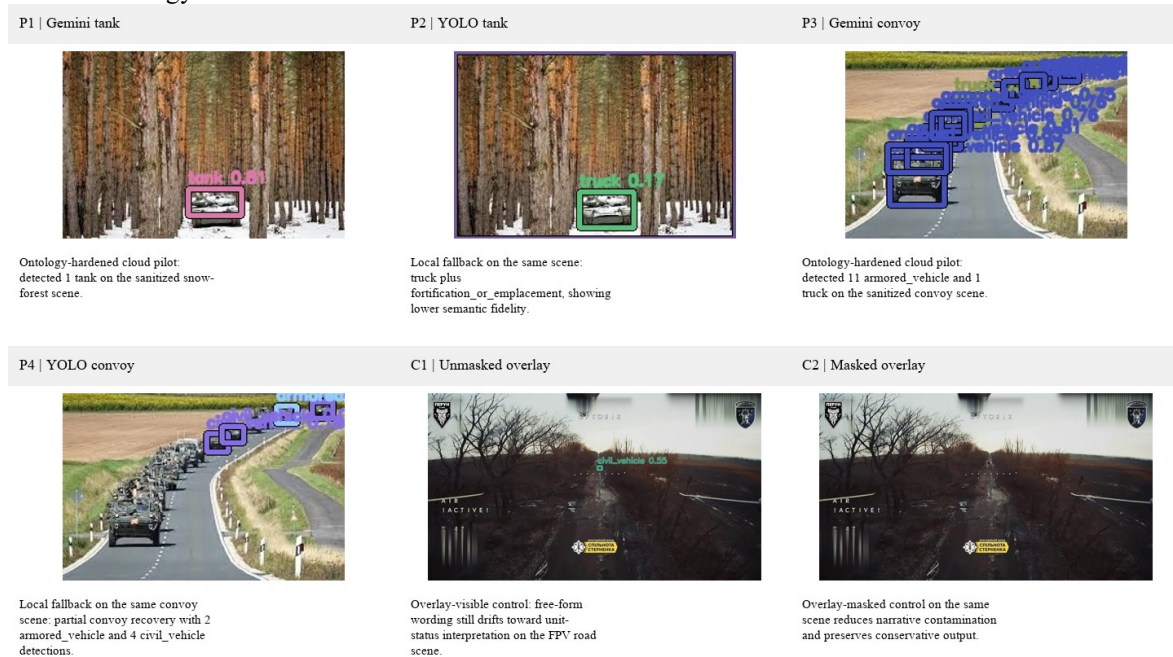


Figure 6. Curated qualitative comparison cases from sanitized pilot scenes and overlay-control conditions.

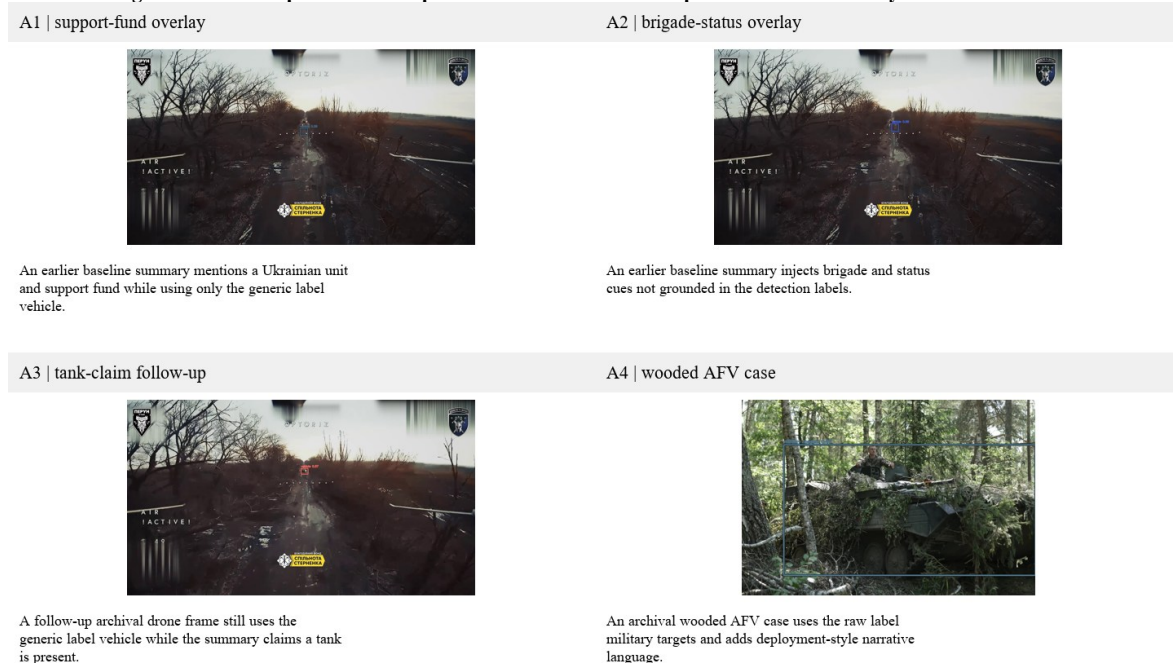


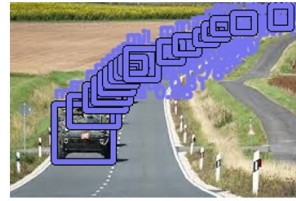
Figure 7. First set of archival qualitative comparison cases illustrating unsupported description and vocabulary drift.

A5 | markdown AFV case



An archival wooded AFV case mixes personnel with military targets and emits markdown-style structured summary text.

A6 | convoy mil-targets



An archival convoy case repeats the raw label mil targets across a dense vehicle column.

A7 | snow target case



An archival snow-forest case uses the raw label military target for a single tank scene.

A8 | snow tank reference



A related snow-forest comparison case returns tank directly, showing how raw vocabulary varied across earlier outputs.

Figure 8. Second set of archival qualitative comparison cases from earlier unconstrained outputs.

ADDITIONAL INFORMATION

AUTHOR CONTRIBUTIONS

Conceptualization, M.H.; methodology, V.T.; validation, M.H.; formal analysis, V.T.; investigation, V.T.; data curation, M.H.; writing-original draft preparation, V.T.; writing-review and editing, M.H.; visualization, V.T.; supervision, M.H.; project administration, M.H. All authors have read and agreed to the published version of the manuscript.

DECLARATION ON THE USE OF GENERATIVE ARTIFICIAL INTELLIGENCE TOOLS

In preparing this work, the author used DeepL Translate and Grammarly for: grammar and spelling checks, paraphrasing, and rephrasing. After using these tools/services, the author reviewed and edited the content and takes full responsibility for the content of this publication.

REFERENCES

- [1] Marcu A., Licaret V., Costea D., Leordeanu M. Semantics through Time: Semi-Supervised Segmentation of Aerial Videos with Iterative Label Propagation. Proceedings of the Asian Conference on Computer Vision. 2020. DOI: 10.1007/978-3-030-69525-5_32. URL: https://openaccess.thecvf.com/content/ACCV2020/html/Marcu_Semantics_through_Time_Semi-supervised_Segmentation_of_Aerial_Videos_with_Iterative_ACCV_2020_paper.html
- [2] Ellsäßer F. J., Nikuze A. Towards a FAIR Metadata Framework for Drone and Uncrewed Aerial Vehicle Data. Scientific Data. 2026. Vol. 13, Art. 57. DOI: 10.1038/s41597-025-06376-9. URL: <https://doi.org/10.1038/s41597-025-06376-9>
- [3] Cheng T., Song L., Ge Y., Liu W., Wang X., Shan Y. YOLO-World: Real-Time Open-Vocabulary Object Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. P. 16901-16911. URL: https://openaccess.thecvf.com/content/CVPR2024/html/Cheng_YOLO-World_Real-Time_Open-Vocabulary_Object_Detection_CVPR_2024_paper.html
- [4] Zhou Y., Li J., Ou C., Yan D., Zhang H., Xue X. UAV-OVD: Open-Vocabulary Object Detection in UAV Imagery via Multi-Level Text-Guided Decoding. Drones. 2025. Vol. 9, No. 7. Art. 495. DOI: 10.3390/drones9070495. URL: <https://www.mdpi.com/2504-446X/9/7/495>
- [5] Zhou Y., Li J., Ou C., Yan D., Zhang H., Xue X. Open-Vocabulary Object Detection in UAV Imagery: A Review and Future Perspectives. Drones. 2025. Vol. 9, No. 8. Art. 557. DOI: 10.3390/drones9080557. URL: <https://www.mdpi.com/2504-446X/9/8/557>
- [6] Xia G.-S., Bai X., Ding J., et al. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018. P. 3974-3983. DOI: 10.1109/CVPR.2018.00418. URL: https://openaccess.thecvf.com/content_cvpr_2018/html/Xia_DOTA_A_Large-Scale_CVPR_2018_paper.html

- [7] Du D., Qi Y., Yu H., et al. The Unmanned Aerial Vehicle Benchmark: Object Detection and Tracking. Proceedings of the European Conference on Computer Vision. 2018. P. 370-386. URL: https://openaccess.thecvf.com/content_ECCV_2018/html/Dawei_Du_The_Unmanned_Aerial_ECCV_2018_paper.html
- [8] Zhu P., Wen L., Bian X., Ling H., Hu Q. Vision Meets Drones: A Challenge. arXiv preprint arXiv:1804.07437. 2018. URL: <https://arxiv.org/abs/1804.07437>
- [9] Basak D., Bhatt S., Kanduri S., Desarkar M. S. Aerial Mirage: Unmasking Hallucinations in Large Vision Language Models. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2025. P. 5500-5508. DOI: 10.1109/WACV61041.2025.00537. URL: https://openaccess.thecvf.com/content/WACV2025/html/Basak_Aerial_Mirage_Unmasking_Hallucinations_in_Large_Vision_Language_Models_WACV_2025_paper.html
- [10] Agrawal A., KV G., Aralikatti R., et al. Towards Mitigating Hallucinations in Large Vision-Language Models by Refining Textual Embeddings. Findings of the Association for Computational Linguistics: ACL 2026. 2026. P. 42032-42048. DOI: 10.18653/v1/2026.findings-acl.2086. URL: <https://aclanthology.org/2026.findings-acl.2086/>
- [11] Cui Y., Qi X., Geng T., Zhang Y., Han S., Guo Y. When Text Hijacks Vision: Benchmarking and Mitigating Text Overlay-Induced Hallucination in Vision Language Models. arXiv preprint arXiv:2604.17375. 2026. URL: <https://arxiv.org/abs/2604.17375>
- [12] Guan T., Liu F., Wu X., et al. HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. P. 14375-14385. DOI: 10.1109/CVPR52733.2024.01363. URL: https://openaccess.thecvf.com/content/CVPR2024/html/Guan_HallusionBench_An_Advanced_Diagnostic_Suite_for_Entangled_Language_Hallucination_and_CVPR_2024_paper.html
- [13] Li Y., Du Y., Zhou K., Wang J., Zhao X., Wen J.-R. Evaluating Object Hallucination in Large Vision-Language Models. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023. P. 292-305. DOI: 10.18653/v1/2023.emnlp-main.20. URL: <https://aclanthology.org/2023.emnlp-main.20/>
- [14] Li C., Im E. W., Fazli P. VidHalluc: Evaluating Temporal Hallucinations in Multimodal Large Language Models for Video Understanding. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2025. P. 13723-13733. URL: https://openaccess.thecvf.com/content/CVPR2025/html/Li_VidHalluc_Evaluating_Temporal_Hallucination_in_Multimodal_Large_Language_Models_for_CVPR_2025_paper.html
- [15] Kraus M., Azimi S. M., Ercelik E., Bahmanyar R., Reinartz P., Knoll A. AerialMPTNet: Multi-Pedestrian Tracking in Aerial Imagery Using Temporal and Graphical Features. 25th International Conference on Pattern Recognition. 2021. P. 2454-2461.
- [16] German Aerospace Center (DLR). EAGLE. URL: <https://www.dlr.de/en/eoc/about-us/remote-sensing-technology-institute/photogrammetry-and-image-analysis/public-datasets/eagle> (accessed: 13.03.2026).
- [17] Li W., Huang Z., Li H., Lu L., Lu Y., Tian X., Shen X., Ye J. Visual Evidence Prompting Mitigates Hallucinations in Large Vision-Language Models. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025. P. 4048-4080. DOI: 10.18653/v1/2025.acl-long.205. URL: <https://aclanthology.org/2025.acl-long.205/>
- [18] Yang L., Zheng Z., Chen B., Zhao Z., Lin C., Shen C. Nullu: Mitigating Object Hallucinations in Large Vision-Language Models via HalluSpace Projection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2025. P. 14635-14645. URL: https://openaccess.thecvf.com/content/CVPR2025/html/Yang_Nullu_Mitigating_Object_Hallucinations_in_Large_Vision-Language_Models_via_HalluSpace_CVPR_2025_paper.html
- [19] Zhao Y., Lv W., Xu S., et al. DETRs Beat YOLOs on Real-time Object Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. P. 16965-16974. DOI: 10.1109/CVPR52733.2024.01605. URL: https://openaccess.thecvf.com/content/CVPR2024/html/Zhao_DETRs_Beat_YOLOs_on_Real-time_Object_Detection_CVPR_2024_paper.html
- [20] Liu Y.-C., Ma C.-Y., He Z., et al. Unbiased Teacher for Semi-Supervised Object Detection. International Conference on Learning Representations. 2021. URL: https://openreview.net/forum?id=MJIve1zgR_
- [21] Xu M., Zhang Z., Hu H., et al. End-to-End Semi-Supervised Object Detection With Soft Teacher. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021. P. 3060-3069. URL: https://openaccess.thecvf.com/content/ICCV2021/html/Xu_End-to-End_Semi-Supervised_Object_Detection_With_Soft_Teacher_ICCV_2021_paper.html
- [22] Mi P., Lin J., Zhou Y., et al. Active Teacher for Semi-Supervised Object Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022. P. 14482-14491. DOI: 10.1109/CVPR52688.2022.01408. URL: https://openaccess.thecvf.com/content/CVPR2022/html/Mi_Active_Teacher_for_Semi-Supervised_Object_Detection_CVPR_2022_paper.html

Supervised_Object_Detection_CVPR_2022_paper.html
[23] Redmon J., Divvala S., Girshick R., Farhadi A. You Only Look Once: Unified, Real-Time Object Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. P. 779-788. DOI: 10.1109/CVPR.2016.91. URL: https://openaccess.thecvf.com/content_cvpr_2016/html/Redmon_You_Only_Look_CVPR_2016_paper.html
[24] Lin T.-Y., Maire M., Belongie S., et al. Microsoft COCO: Common Objects in Context. European Conference on Computer Vision. 2014. P. 740-755. DOI: 10.1007/978-3-319-10602-1_48. URL: <https://arxiv.org/abs/1405.0312>
[25] Kharlashkin L., Morooka E. V., Tereshchenko Y., Hämmäläinen M. ORACLE: Time-Dependent Recursive Summary Graphs for Foresight on News Data Using LLMs. Proceedings of the 10th International Workshop on Computational Linguistics for Uralic Languages. 2025. P. 72-76. URL: <https://aclanthology.org/2025.iwclul-1.10/>
[26] German Aerospace Center (DLR). AerialMPT: A Dataset for Pedestrian Tracking in Aerial Imagery.

URL: <https://www.dlr.de/en/eoc/about-us/remote-sensing-technology-institute/photogrammetry-and-image-analysis/public-datasets/aerialmpt-a-dataset-for-pedestrian-tracking-in-aerial-imagery> (accessed: 13.03.2026).
[27] German Aerospace Center (DLR). DLR-MVDA. URL: <https://www.dlr.de/en/eoc/about-us/remote-sensing-technology-institute/photogrammetry-and-image-analysis/public-datasets/dlr-mvda> (accessed: 13.03.2026).
[28] German Aerospace Center (DLR). DLR's Aerial Crowd Dataset (DLR-ACD). URL: <https://www.dlr.de/en/eoc/about-us/remote-sensing-technology-institute/photogrammetry-and-image-analysis/public-datasets/dlr-acd> (accessed: 13.03.2026).
[29] Tereshchenko Y., Hämmäläinen M. A Comparative Analysis of Ethical and Safety Gaps in LLMs using Relative Danger Coefficient. Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities. 2025. P. 464-477. DOI: 10.18653/v1/2025.nlp4dh-1.40. URL: <https://aclanthology.org/2025.nlp4dh-1.40/>

Єгор ТЕРЕЩЕНКО, Міка ХЯМЯЛЯЙНЕН
Метрополія Університет прикладних наук, Гельсінкі, Фінляндія

ГАЛЮЦИНАЦІЙНО-СТІЙКЕ ФОРМУВАННЯ UAV-ДАТАСЕТУ ДЛЯ ISR

У статті досліджено галюцинаційно-стійке анотування БПЛА-подібних зображень у публікаційно-безпечній постановці та розглянуто його як задачу вимірювання, надійності й розгортання, а не як коротку демонстрацію на основі промптового інжинірингу. Запропонована схема поєднує нормалізацію онтології, маскуванню текстових накладок, режим обґрунтованих підсумків, безпечну обробку медіа з Unicode, контроль профілів підказок і структуровану фіксацію результатів, що дає змогу спільно аудіювати мітки, підрахунки та текстові описи. Для усунення розриву між зручними мультимодальними демонстраціями та науково придатним доказом робота поєднує архівний аудит 7 історичних кадрів військового типу з публічним проксі-бенчмарком обсягом 350 аерознімків і кадрів із наборів AerialMPT, DLR-MVDA та DLR-ACD. Виконаний публічний тест містить 274,581 еталонних об'єктів, що переводить емпіричну основу роботи від вузького пілотного дослідження до оцінювання на сотнях прикладів. Локальна базова лінія YOLO-World завершила обробку всіх 350 публічних зразків із середньою затримкою 0.1807 с на зображення, але суттєво недооцінювала щільні сцени: середня абсолютна помилка підрахунку становила 143.889 для AerialMPT, 297.300 для повнорозмірних сцен DLR-MVDA та 6848.242 для DLR-ACD. Для узгодженого хмарно-локального порівняння в роботі введено цільний фрагментний тест DLR-MVDA із 10 похідними кроками 1024x1024 та 315 розміченими транспортними засобами. Трипрофільне дослідження абляції Gemini Robotics-ER показало, що конструкція підказок матеріально змінює якість результату: загальний профіль завершив усі фрагменти, але сильно переоцінював кількість об'єктів, обмежений ISR-профіль зменшив похибку, але втратив один запит, а найкращий профіль, coarse-to-fine, досяг MAE 5.300, MAPE 10.161% та завершення 10 / 10. Додаткове відеодослідження на основі публічно доступних тактичних кадрів ізолює контроль тексту: вільна умова дала вихід за межі замкненої онтології у 21 з 21 непорожніх підсумків, тоді як узгоджена обґрунтована умова зменшила це значення до 0 з 20 та скоротила затримку з 491.070 с до 272.711 с. Архівний аудит додатково виявив невідтримувані формулювання про підрозділ, статус або намір у 4 із 7 історичних кадрів (57.1%) та чотири сирі варіанти міток (mil targets, military target, military targets, vehicle) для однієї концепції броньованої платформи. Отже, внеском статті є експериментально обґрунтований підхід до галюцинаційно-стійкого формування БПЛА-датасетів, у якому спільно оцінюються семантичний контроль, якість підрахунку, чутливість до профілю підказок та надійність виконання.

Ключові слова: БПЛА ISR, схема анотування, grounded-summary режим, публічний аеробенчмарк, open-vocabulary детекція, локальне безпечне розгортання