UDC 004.93:623.746-519 DOI:10.34229/2707-451X.26.2.6

GEN-THERMAL-UAV: A SYNTHETIC BENCHMARK FOR MULTISCALE FIXED-WING UAV TRACKING IN THERMAL INFRARED VIDEO

Yaroslav Smolin *, Viacheslav Liskin

ORCIDs: 0009-0007-9744-6496; 0000-0002-9418-0633

National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
** Correspondence: smolin.yaroslav@lll.kpi.ua*

УДК 004.93:623.746-519

Gen-Thermal-UAV: синтетичний бенчмарк для відстеження БпЛА з фіксованим крилом у тепловому інфрачервоному відео

Я.С. Смолін *, В.О. Ліскін

Національний технічний університет України «КПІ імені Ігоря Сікорського», Київ
** Листування: smolin.yaroslav@lll.kpi.ua*

Вступ. Масове поширення безпілотних літальних апаратів (БпЛА) літакового типу, зокрема баражуючих боєприпасів, створює суттєві виклики для безпеки повітряного простору. Їх виявлення значною мірою ґрунтується на використанні теплових інфрачервоних камер, які забезпечують пасивний моніторинг у режимі 24/7. Ключовою проблемою розробки таких систем є задача «мульти-масштабного наближення»: супровід цілі в процесі швидкого переходу від далекого субпіксельного «точкового» відбитку до близького, добре розрізненого об'єкта. Цей перехід є критичною точкою відмови сучасних систем проти повітряної оборони (ППО) ближньої дії, відомою як проблема передавання супроводу. Існуючі набори даних не відображають цієї безперервної еволюції через небезпеку та високу вартість зйомки реальних повітряних зближень за курсом зіткнення. Такий «вакуум» даних гальмує розвиток надійних алгоритмів протидії БпЛА, оскільки традиційні синтетичні дані часто не відображають термодинамічну достовірність реальних сенсорів.

Мета роботи полягає у створені Gen-Thermal-UAV – нового синтетичного набору даних, розробленого для заповнення цієї прогалини, а також у пропозиції методології, що використовує сучасні відеодіфузійні моделі (Gemini Veo 3) на основі декількох реальних опорних кадрів. Такий підхід спрямований на генерацію високодостовірних синтетичних відео, які зберігають автентичні характеристики сенсорів і водночас моделюють різноманітні траєкторії польоту, забезпечуючи навчання моделей детектування та трекінгу, стійких до екстремальних змін масштабу.

Методологія. Застосовано генеративний AI-конвеєр, керований опорними кадрами. Замість повної генерації даних «з нуля» використано підхід «зображення – відео», прив'язаний до двох реальних тепловізійних зображень: дальнього (розмитої точки) та ближнього (розрізненої літакаподібної цілі). Такий підхід забезпечує термодинамічну достовірність, оскільки дифузійна модель поширює реальний шум сенсора, розмиття та теплові сигнатури, присутні в стартових зображеннях, уздовж реалістичних траєкторій польоту. Модель дифузії передбачає рух розподілу пікселів, фактично «галюцинуючи» збереження фізики, а не спираючись на низько якісну растеризацію. Розроблено структуровану таксономію промптів для генеративної моделі, що дає різноманітність та узгодженість сценаріїв. Отримані відео анотовано напівавтоматично за допомогою Segment Anything Model 2 (SAM 2), яка використовує часову узгодженість для zero-shot-розмітки; результати фільтрувалися за порогом довіри 85 %.

Результати. Набір даних Gen-Thermal-UAV містить 220 відео (1 760 секунд, близько 42 000 кадрів) із роздільною здатністю 720p, що відображають сценарії «повітря – повітря» для взаємодії літакового типу БпЛА. Це перший набір даних, який фіксує безперервний перехід від точки до об'єкта в тепловізійному діапазоні. Порівняльний аналіз підтверджує його унікальне положення на перетині теплової модальності, платформи «повітря – повітря» та екстремальної мульти-масштабної динаміки. Це відрізняє його від таких еталонів, як AOT, HIT-UAV та Anti-UAV410.

Висновки. Запропонований підхід Gen-Thermal-UAV усуває важливу прогалину в галузі комп'ютерного зору, пов'язану із застосувань протидії БпЛА. Розроблена та верифікована методологія демонструє, що генеративний ШІ, обмежений реальними опорними кадрами, здатний створювати фізично узгоджені навчальні дані для небезпечних або рідкісних сценаріїв. Представлене дослідження не лише пропонує критично важливий еталон, але й формує відтворюваний протокол генерації та авто-розмітки синтетичних даних, демократизуючи доступ до навчання на «крайових» випадках та сприяючи швидкій адаптації до нових загроз.

Ключові слова: тепловізійне інфрачервоне стеження, безпілотні літальні апарати, генерація синтетичних даних, відеодифузійні моделі, мульти-масштабне виявлення, комп'ютерний зір.

Abstract

Introduction. The proliferation of fixed-wing Unmanned Aerial Vehicles (UAVs), such as loitering munitions, presents a significant challenge to airspace security. Detection relies heavily on Thermal Infrared (TIR) imaging for 24/7 passive monitoring. A critical challenge in developing these systems is the "Multiscale Approach" problem: tracking a target as it rapidly transitions from a distant sub-pixel dot to a close-range resolved object. This transition is a critical failure point for modern defense systems, known as the "handover" problem in SHORAD. Existing datasets fail to capture this continuous evolution due to the dangers and costs associated with filming air-to-air collision courses. This data vacuum hinders the development of robust Counter-UAS (C-UAS) algorithms, as traditional synthetic data often lacks the thermodynamic fidelity of real sensors.

The purpose of the paper is to introduce Gen-Thermal-UAV, a novel synthetic dataset designed to fill this gap, and to propose a "Seed-Driven" methodology utilizing advanced video diffusion models (Gemini Veo 3). This approach aims to generate high-fidelity synthetic videos that maintain authentic sensor characteristics while simulating diverse flight trajectories, enabling the training of end-to-end trackers robust to extreme scale changes.

Methodology. We employed a Seed-Driven Generative AI pipeline. Instead of generating data from scratch, we used Image-to-Video generation anchored by two real thermal images: one far-field (a blurry dot) and one near-field (a resolved plane). This approach ensures thermodynamic fidelity, as the diffusion model propagates the real sensor noise, blur, and heat signatures present in the seed images along realistic flight paths. The diffusion model predicts the motion of the pixel distribution, effectively "hallucinating" the preservation of physics rather than relying on low-fidelity rasterization. A structured prompt engineering taxonomy was developed to constrain the generative model to scientific consistency. The resulting videos were automatically annotated using the Segment Anything Model 2 (SAM 2), leveraging temporal consistency for zero-shot labeling, validated by an 85% confidence filter.

Results. Gen-Thermal-UAV comprises 220 videos (1,760 seconds, approx. 42,000 frames) at 720p resolution, depicting air-to-air fixed-wing engagement scenarios. It is the first dataset to capture the continuous dot-to-object transition in the thermal domain. A comparative analysis confirms its unique position at the intersection of Thermal modality, Air-to-Air platform, and Extreme multiscale dynamics, distinguishing it from benchmarks like AOT, HIT-UAV, and Anti-UAV410.

Conclusions. Gen-Thermal-UAV addresses a high-value gap in the computer vision landscape for C-UAS applications. The verified methodology demonstrates that Generative AI, when constrained by real-world seed data, can produce physics-compliant training data for dangerous or rare scenarios. This work not only provides a crucial benchmark but also establishes a reproducible protocol for generating and auto-labeling synthetic data, democratizing access to "edge case" training and facilitating rapid adaptation to emerging threats.

Keywords: thermal infrared tracking, unmanned aerial vehicles, synthetic data generation, video diffusion models, multiscale detection, computer vision.

Introduction. The contemporary geopolitical landscape has been fundamentally reshaped by the proliferation of Unmanned Aerial Systems (UAS), with a decisive shift toward fixed-wing platforms such as loitering munitions and high-altitude surveillance assets. These systems, characterized by high speeds, high-altitude operations, and complex kinematics, present a challenge to airspace security. Detection of these threats increasingly depends on Thermal Infrared (TIR) imaging, which offers passive, 24/7 monitoring capabilities by detecting thermal emissivity against the cold background of the sky, circumventing limitations of radar and visible-spectrum sensors.

However, the efficacy of TIR-based defense systems is critically undermined by a lack of representative training data for Computer Vision algorithms. The central problem is the Multiscale Approach paradox. A robust system must identify a threat at maximum range, when the target is an almost point-like "thermal dot" (the domain of Infrared Small Target Detection, IRSTD), and maintain a continuous lock as it closes the distance and rapidly evolves into a resolved object with discernible geometric features.

This transition represents a critical failure mode in modern Short-Range Air Defense (SHORAD). In a typical "system of systems" architecture, a radar detects a target, and a thermal camera slews to identify it. The critical point of failure is often the handover. The tracker must lock onto the initial dot and maintain the lock as the target approaches and changes shape. If the tracker fails during this multiscale transition, the engagement fails. Current algorithms struggle with this handover: models trained on IRSTD datasets fail once the target develops internal texture, while standard trackers fail to initialize on featureless dots.

Bridging this gap requires data capturing the continuous temporal evolution of an air-to-air interception. Capturing such data in the real world is logistically prohibitive due to the risks and costs of collision-course flights. Consequently, the research community relies on fragmented datasets, each covering only a subset of the sensing modality, platform geometry, or scale dynamics relevant to fixed-wing interception.

This paper introduces Gen-Thermal-UAV, a research contribution addressing this specific lacuna through Generative Artificial Intelligence. By leveraging the Gemini Veo 3 video diffusion model [1], anchored by real thermal "seed" images, we generate a dataset that combines the thermodynamic realism of authentic sensor data with the trajectory diversity of simulation. This approach enables the training of end-to-end trackers that learn a unified feature representation robust to extreme scale changes, directly addressing the SHORAD handover problem and potentially increasing the kill probability (*Pk*) of automated defense systems.

Furthermore, this seed-driven methodology democratizes the creation of datasets for "Black Swan" events – rare but catastrophic scenarios such as head-on collisions that cannot be safely filmed. A researcher needs only one image of a new threat (e.g., a grainy thermal capture of a novel enemy drone) to generate thousands of synthetic training videos. This allows for rapid adaptation of defensive algorithms to new threats, reducing the "time to adaptation" from months to hours.

This work presents three primary contributions: (1) the Gen-Thermal-UAV dataset (220 videos) of synthetic thermal air-to-air interception scenarios for fixed-wing UAVs; (2) a seed-driven generation pipeline that distills the thermodynamic properties of real thermal images into extensive video data via a state-of-the-art video diffusion model [1]; and (3) verification of this dataset as a benchmark targeting the continuous dot-to-object transition, validated by an automated zero-shot labeling workflow using the Segment Anything Model 2 (SAM 2) [2].

Analysis of Recent Studies and Publications. To verify the novelty of Gen-Thermal-UAV, we analyze the existing landscape of UAV detection and tracking datasets. The current ecosystem is fragmented across sensing modalities, platforms, and target definitions, leaving a specific gap for thermal fixed-wing air-to-air engagement.

The most prominent thermal benchmarks are the Anti-UAV series, including the recently proposed Anti-UAV410 dataset [3]. Anti-UAV410 provides high-quality thermal sequences of drones in challenging outdoor scenes but predominantly features rotary-wing quadcopters, whose flight dynamics (hovering, slow translation) differ fundamentally from the high-speed, high-inertia trajectories of fixed-wing aircraft. Moreover, sensing is typically ground-to-air, precluding truly dynamic air-to-air backgrounds with parallax-rich cloud motion. Safety constraints further prevent the inclusion of true collision-course trajectories that would capture the radial expansion of an incoming fixed-wing threat.

In the aerial domain, the Airborne Object Tracking (AOT) dataset [4] offers massive scale for air-to-air tracking, with millions of annotated frames collected from aircraft platforms. However, AOT is exclusively in the RGB/gray domain; transferring learning from RGB to thermal imagery is non-trivial due to the un-

derlying physics mismatch: RGB sensors observe reflected light, while TIR sensors measure emitted thermal radiation. AOT cannot teach a network to interpret sensor-specific noise statistics, blooming, or scintillation patterns that are critical for robust thermal tracking.

The HIT-UAV dataset [5] provides high-altitude thermal data, focusing on air-to-ground surveillance. Targets include vehicles and pedestrians imaged against hot earth backgrounds, with clutter statistics dominated by ground emissivity. These statistics are incompatible with the "cold sky / hot aircraft" regime of anti-aircraft tracking. The camera geometry is also different: steep look-down angles in HIT-UAV versus near-horizon sky backgrounds in air-to-air engagements.

Beyond C-UAV-centric benchmarks, BIRDSAI [6] offers an aerial TIR dataset collected from a fixed-wing UAV for wildlife conservation. It contains real and synthetic thermal videos of humans and animals under poaching scenarios and demonstrates the value of combining real and simulator-generated thermal data. However, BIRDSAI focuses on ground targets and does not model the extreme z-axis scale change associated with head-on fixed-wing interception.

Datasets focusing on Infrared Small Target Detection (IRSTD), such as NUAA-SIRST [7], NUDT-SIRST [8], and IRSTD-1K [9], treat the task primarily as binary segmentation on static images. They provide high-quality labeled data for dot-like targets but lack temporal coherence and do not cover the transition from sub-pixel anomalies to fully resolved aircraft. IRSTD algorithms trained on these datasets often fail once the target grows larger than a few pixels and develops internal texture.

To mitigate data scarcity in IRSTD, several works have explored synthetic infrared generation. A recent study proposed a GAN-based method for infrared dim and small-target sequence dataset generation, showing that synthetic data can enhance deep IRSTD models when carefully designed [10]. Nonetheless, many GAN- and game-engine–based methods produce sky-background sequences without realistic sensor artifacts or complex 3D air-to-air kinematics.

Beyond the C-UAV datasets above, the broader UAV tracking community has released large-scale RGB and RGB-T benchmarks. UAV123 [11] introduces 123 low-altitude aerial sequences (more than 110k frames) shot from a UAV platform, together with a photorealistic simulator tailored for aerial tracking. VTUAV (Visible–Thermal UAV Tracking) [12] provides 500 paired RGB-T sequences with 1.7 million high-resolution (1920×1080) frame pairs and supports short-term, long-term and segmentation-mask tracking evaluation. VisDrone [13] offers over 260 video clips and more than 10k images captured by drones over 14 cities, with tasks including image/video detection and single/multi-object tracking. These datasets are invaluable for evaluating tracking architectures and multi-modal fusion, but they focus on ground targets or larger, clearly resolved UAVs and predominantly operate in RGB or RGB-T modalities rather than pure TIR air-to-air engagements.

For drone detection itself, Svanström et al. compiled a multi-sensor drone detection dataset that includes synchronized infrared, visible and audio streams recorded at several airports [14]. The database covers multiple drones and confuser classes such as birds, airplanes and helicopters and is explicitly organized by range bands following the Johnson DRI criteria. While highly relevant for C-UAV research, acquisition ranges are constrained by visual-line-of-sight regulations (~200 m), and the geometry remains ground-to-air rather than air-to-air.

On the algorithmic side, recent surveys on infrared small-target segmentation and detection networks provide comprehensive taxonomies of IRSTD architectures, losses and publicly available datasets, and they unanimously identify data scarcity, temporal incoherence and limited scenario diversity as key bottlenecks [15, 16]. However, these works mostly discuss static or quasi-static target scales and do not supply thermal air-to-air video of fixed-wing aircraft.

A new wave of TIR anti-UAV benchmarks has also appeared. CST Anti-UAV [17] is a thermal dataset of 220 sequences with tiny UAVs in highly cluttered scenes and over 240k manual bounding-box annotations, designed for single-object tracking in complex environments. It exposes pronounced failure modes of current trackers on tiny targets. At the method level, CAMTracker introduces a contrastive-augmented

memory network that significantly improves long-term anti-UAV tracking performance on Anti-UAV benchmarks in TIR videos [18].

As summarized in Table 1, there is still no publicly available dataset that simultaneously offers: (i) thermal sensing, (ii) air-to-air sensor geometry, and (iii) continuous extreme dot-to-object scale changes for fixed-wing UAVs. Existing TIR benchmarks focus on drones observed from ground-based sensors or on static/dot targets; large RGB and RGB-T benchmarks and multi-sensor datasets operate in different sensing regimes and geometries. Gen-Thermal-UAV is explicitly designed to fill this gap.

Summary of sensing modality, platform geometry, target type, dataset size, and effective scale regime. Gen-Thermal-UAV is the only dataset that simultaneously provides thermal sensing, air-to-air geometry, and extreme dot-to-object scale transitions for fixed-wing UAVs.

TABLE 1. Comparison of existing UAV / IR small-target datasets with the proposed Gen-Thermal-UAV benchmark

Dataset	Modality	Platform	Target type	Videos/images	Scale
AOT [4]	RGB/Gray	Air-to-Air	Planes/rotors	4,943/5.9M	Low (long range)
HIT-UAV [5]	Thermal	Air-to-Ground	ground objects	– /2,898	Low (top-down)
Anti-UAV410 [3]	Thermal	Ground-to-Air	Rotary (mostly)	410/438k	Mixed
CST Anti-UAV [17]	Thermal	Ground-to-Air	Tiny UAVs	220/240k	Mostly tiny-scale
BIRDSAI [6]	Thermal (real+sim)	Air-to-Ground	ground objects	48/50k	Low–medium (ground)
NUAA-SIRST [7]	Thermal	Mixed	Small/dim targets	– /400	Static (dot-level)
NUDT-SIRST [8]	Thermal	Mixed	Small/dim targets	– /1k	Static (dot-level)
IRSTD-1K [9]	Thermal	Mixed	Small/dim targets	– /1k	Static (dot-level)
GAN-SIRST [10]	Thermal (synthetic)	Simulated	Small/dim targets	– /1,327	Static / small-motion
UAV123 [11]	RGB	Air-to-Ground	ground objects	123/110k	Medium (viewpoint)
VTUAV [12]	RGB-T	Air-to-Ground	ground objects	500/1.7M	Medium (RGB-T)
VisDrone [13]	RGB	Air-to-Ground	ground objects	263/179k	Medium (crowded scenes)
Multi-sensor [14]	RGB+Thermal+Audio	Ground-to-Air	Drones/confusers	450/205k	Short-range bands
Gen-Thermal-UAV	Thermal (synthetic)	Air-to-Air	Fixed-wing	220/42k	Extreme (dot↔object)

Methodology. The proposed methodology employs a seed-driven Generative AI pipeline. We treat generation as a form of "knowledge distillation", where the thermodynamic truth of a real thermal image is expanded into a temporal sequence by a video diffusion model.

We use image-to-video generation via the Gemini Veo 3 model [1], anchored by real thermal seeds.

Diffusion models are probabilistic generative models that learn the data distribution p(x) by reversing a gradual noise-addition process. When we provide a real thermal seed, we supply a sample x0 drawn from the true distribution of thermal physics. The seed inherently contains atmospheric blur, sensor grain, bloo ming and thermal contrast that conventional game engines fail to simulate accurately.

The diffusion model's task is not to rasterize the next frame from geometric primitives, but to predict the motion of the existing pixel distribution. It infers how a noisy, blurry patch of pixels (the UAV) evolves over time within a noisy background. Because the model has been trained on vast quantities of real videos, it implicitly encodes the statistics of motion (optical flow) and propagates the texture of the seed – the real thermal physics – along plausible flight paths. This effectively "hallucinates" physics-preserving motion, producing videos that retain subtle effects such as blooming and scintillation present in the original seed.

We utilize two distinct seeds: Seed A (Fig. 1, *a*), a real thermal capture of a fixed-wing UAV at extreme distance (>700m), and Seed B, a real thermal capture at close range, near 50 m (Fig. 1, *b*). Seed A generates 81 videos emphasizing detection and approach under low signal-to-noise ratio (SNR), while Seed B generates 139 videos focusing on chasing and maneuvering, including high-G turns and evasive behavior. Together, the seeds span the full multiscale continuum from dot-like detection to high-detail terminal tracking.

a *b*

FIG. 1. Real thermal seed images used for video diffusion: *a* – Seed A – far-field IRSTD-scale fixed-wing UAV at 700 m range, appearing as a blurry thermal dot; *b* – Seed B – near-field fixed-wing UAV at <50 m, with a resolved airframe and clearly visible hot engine region. These two seeds anchor the multiscale dot-to-object continuum modeled in Gen-Thermal-UAV

Structured Prompt Engineering. To achieve scientific consistency rather than artistic randomness, we developed a structured prompt taxonomy (Table 2). This taxonomy constrains the generative model to obey the physical and operational characteristics of fixed-wing flight and TIR imaging.

The components are: (1) Environmental context, defining background and atmosphere; (2) Sensor modality, enforcing the visual style of the hardware; (3) Subject definition, describing target morphology; (4) Scale and evolution, specifying temporal change in apparent size; and (5) Motion dynamics, defining trajectory and camera behavior.

This structured approach acts as a constraint-satisfaction mechanism. By explicitly including negative prompts (e.g., “bad quality”), we encourage the model to generate data that resembles real, imperfect military sensors rather than polished CGI renders, reducing the reality gap.

TABLE 2. Structured prompt taxonomy for seed-driven thermal video synthesis. Prompt components used with Gemini Veo 3 to generate physics-consistent thermal air-to-air videos from real seeds. Each component constrains a different aspect of the simulation: environment, sensor modality, target morphology, multiscale evolution, and motion dynamics

Component	Function	Example
Environmental context	Defines background and atmosphere	“High altitude, clouds, warzone environment.”
Sensor modality	Enforces visual style of hardware	“Infrared footage (white-hot), bad quality, high sensor noise, monochromatic.”
Subject definition	Describes target morphology	“Small fixed-wing plane, engine area is hotter.”
Scale & evolution	Describes temporal change in size	“Tiny dot becomes visible plane” (approach scenario)
Motion dynamics	Defines trajectory and camera behavior	“Camera chase, shaky rapid movement, banking turn.”

Automated Labeling with SAM 2. We automate annotation using SAM 2 [2]. A single bounding box is drawn on the seed image (frame 0). SAM 2 employs a memory-bank architecture and temporal attention to propagate this mask through subsequent frames, producing high-quality segmentation masks over time.

Because the video is generated from a single seed, there is perfect feature continuity at the start, allowing SAM 2 to lock onto the target's thermal gradient reliably. We apply a confidence filter: if SAM 2's internal confidence for the propagated mask drops below 85%, the track is terminated and the remaining frames are discarded. This ensures that only high-confidence, pixel-accurate annotations are retained, yielding approximately 42,000 tight bounding boxes and masks. All resulting tracks were then inspected with a fast manual sanity check to remove obvious failures (missed targets, mask drift, identity switches) and to verify overall annotation consistency across sequences.

The Gen-Thermal-UAV Dataset. The Gen-Thermal-UAV dataset is designed to stress-test trackers on approach and chasing behaviors that are under-represented or missing in current benchmarks.

The dataset contains 220 videos, totaling approximately 1,760 seconds (over 42,000 frames, random 15 frames shown in Fig. 2) of air-to-air engagement. All videos are at 1280 × 720 (720p) resolution and 24 fps. Data are grouped by the input seed: Subset A ("Far-Field" collection, 81 videos) and Subset B ("Near-Field" collection, 139 videos).

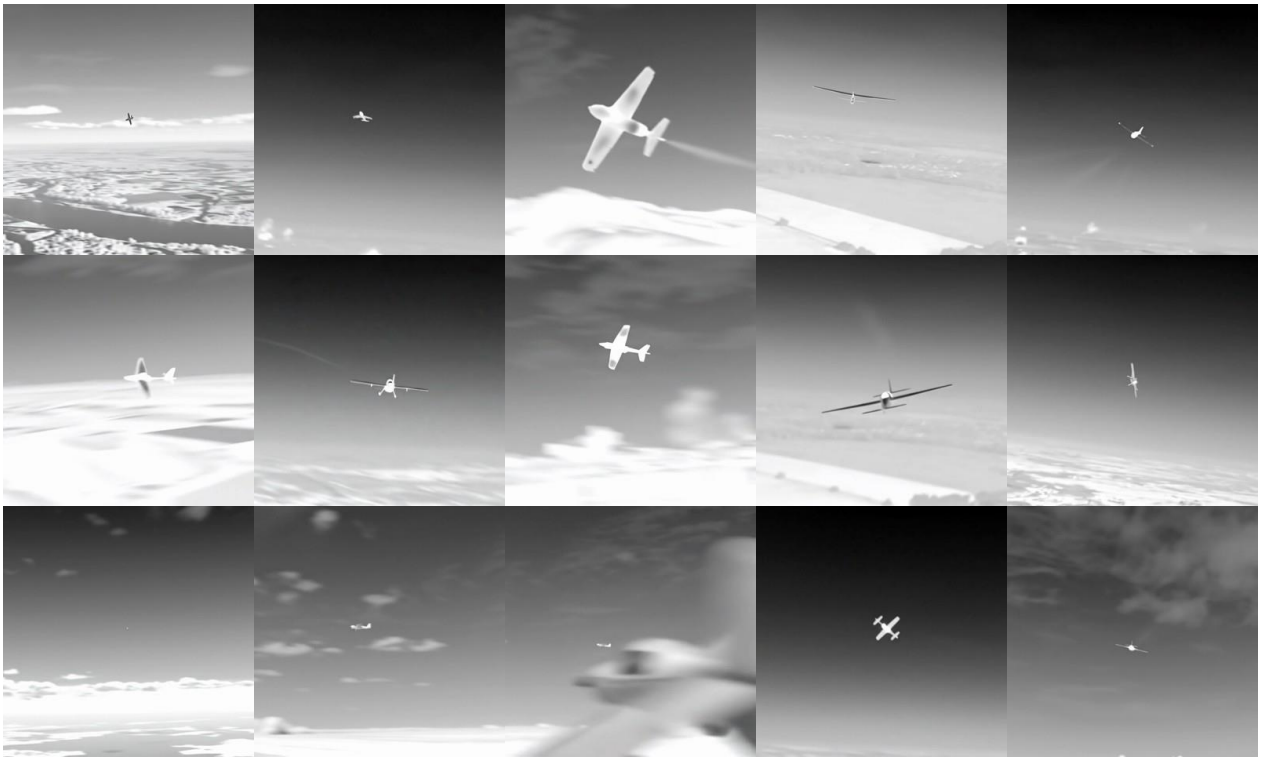

FIG. 2. Representative frames from the Gen-Thermal-UAV dataset. Randomly sampled frames illustrating (i) the dot-scale detection regime, (ii) the continuous dot-to-object expansion as the target approaches, and (iii) close-range maneuvering of fixed-wing UAVs in diverse thermal backgrounds

Each frame is annotated with a bounding box and a binary mask for the target UAV, enabling both tracking and segmentation-based evaluation protocols. These labels are generated automatically by SAM 2 and subsequently verified via a rapid manual pass to sanity-check the tracks and filter residual labeling errors. We assert that Gen-Thermal-UAV is novel based on three pillars of innovation.

First, the "approach vector" novelty: as summarized in Table 1, no existing public dataset provides thermal air-to-air video of fixed-wing UAVs undergoing extreme multiscale changes. Anti-UAV410 [3] and CST Anti-UAV [17] focus on drones observed from ground-based sensors; large aerial benchmarks such as AOT [4], UAV123 [11] and VTUAV [12] operate in RGB or RGB-T modalities; VisDrone [13] and HIT-

UAV [5] emphasize ground targets; BIRDSAI [6] covers aerial TIR data but for conservation scenarios and ground objects; IRSTD benchmarks [7–9] provide static dot-like targets without temporal evolution. Gen-Thermal-UAV is therefore, to the best of our knowledge, the first benchmark to focus explicitly on the continuous thermal "approach" scenario for fixed-wing targets.

Second, the "seed" novelty: our approach verifies a new paradigm of using real thermal seeds inside a video diffusion model. Unlike polygon-based rendering, which tends to produce unrealistic hard edges and over-clean imagery, our method preserves the noise texture, blooming and heat-gradient structure present in operational sensors. The diffusion model "hallucinates" motion of the existing real pixels rather than creating synthetic objects from scratch, thereby maintaining thermodynamic fidelity [1,10].

Third, zero-shot annotation verification: the integration of SAM 2 [2] verifies a novel, scalable workflow that combines the visual fidelity of real data (through seeds) with the labeling efficiency of synthetic data. This zero-shot annotation pipeline eliminates manual bounding-box drawing and mask refinement, reducing the cost of scaling to new sensor modalities or threat classes. The same scheme—seed image + structured prompt + SAM 2 auto-labeling – extends naturally to maritime, ground, or space-based sensing regimes.

Future Work and Discussion. The immediate next step is to establish a benchmark on Gen-Thermal-UAV using state-of-the-art trackers and IRSTD-inspired detectors [3, 5, 7–9, 17, 18]. Recent work such as CAMTracker [18] shows that carefully designed memory and contrastive modules can significantly improve TIR anti-UAV tracking on Anti-UAV benchmarks, yet performance remains limited when targets are tiny and fast-moving. We hypothesize that standard short-term trackers will struggle even more on our dataset, where the apparent scale of the target changes by orders of magnitude within seconds.

A key research question is the efficacy of generative video as data augmentation. We propose an experiment comparing a baseline detector trained only on the two real seed images versus an augmented model trained on the full set of 220 synthetic videos. We hypothesize that the augmented model will show significantly higher average precision on real-world test data, similar to improvements reported in recent anti-UAV work that combines real videos with synthesized UAV data [10,19].

In particular, Liu et al. construct Anti-MUAV15, a multi-object thermal dataset complemented by synthetic UAV sequences, and show that learning an adaptive detection–tracking collaboration on augmented data yields more accurate and robust anti-UAV systems [19]. Our seed-driven diffusion pipeline offers a complementary way to synthesize approach trajectories and could be integrated into similar joint detection–tracking frameworks.

Finally, the pipeline – Seed Image + Structured Prompt + SAM 2 Auto-labeling – is inherently domain-agnostic. This seed-driven approach enables the multiplication of rare or dangerous events across domains such as maritime search and rescue, ground-based surveillance, and space domain awareness. Recent anti-UAV surveys explicitly highlight diffusion-based data synthesis and multi-modal fusion as key future directions for robust UAV perception [20]. Gen-Thermal-UAV provides a concrete instantiation of this agenda in the thermal air-to-air regime and suggests that similar seed-driven generative benchmarks could be constructed for other sensing geometries and threat models. Furthermore, such high-fidelity tracking data is a critical prerequisite for advanced interception logic, such as the two-level routing algorithms for moving targets proposed in [21], bridging the gap between perception and actuation.

Conclusions. Gen-Thermal-UAV addresses a specific, high-value gap in the counter-UAV (C-UAV) landscape: thermal video of fixed-wing aircraft in dynamic air-to-air engagement scenarios. By providing data that covers the entire approach – from IRSTD-scale dot detection to terminal object tracking – it enables the training of next-generation trackers capable of handling the multiscale transition that currently undermines SHORAD handover.

The verification of our methodology demonstrates that Generative AI, when constrained by real-world seed data, can bridge the gap between static image datasets and dynamic video requirements. Through comparative analysis against existing RGB, RGB-T, thermal and IRSTD benchmarks, we confirm that Gen-

Thermal-UAV occupies a unique niche: synthetic yet physics-preserving TIR air-to-air data with extreme scale dynamics. Beyond this specific dataset, the proposed pipeline offers a reproducible protocol for generating and auto-labeling synthetic data, democratizing access to “edge-case” training scenarios and supporting rapid adaptation to emerging threats.

Authorship contribution: Yaroslav Smolin – investigation, conceptualization, methodology, formal analysis, writing – original draft; Viacheslav Liskin – supervision, conceptualization, methodology, formal analysis, resources, writing – review & editing.

Data availability. The Gen-Thermal-UAV dataset supporting the findings of this study will be made openly available at https://zenodo.org/records/19582012.

Funding. The author(s) received no financial support for the research, authorship and/or publication of this article.

References

1. Google DeepMind. Veo 3 Technical Report. 2025. https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf (accessed: 03.12.2025)
2. Ravi N., Gabeur V., Hu Y-T., Cheng Z., Schick A., Huang P-Y., et al. SAM 2: Segment Anything in Images and Videos. *arXiv*. 2024. arXiv:2408.00714. https://doi.org/10.48550/arXiv.2408.00714 (accessed: 03.12.2025)
3. Huang B., Li J., Chen J., Wang G., Zhao J., Xu T. Anti-UAV410: A Thermal Infrared Benchmark and Customized Scheme for Tracking Drones in the Wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2024. **46** (5). P. 2852–2865. https://doi.org/10.1109/TPAMI.2023.3335338
4. Amazon Web Services. Airborne Object Tracking Dataset. https://registry.opendata.aws/airborne-object-tracking/ (accessed: 03.12.2025)
5. Suo J., Wang T., Zhang X., Chen H., Zhou W., Shi W. HIT-UAV: A high-altitude infrared thermal dataset for unmanned aerial vehicle-based object detection. *Scientific Data*. 2023. **10** (1). P. 227. https://doi.org/10.1038/s41597-023-02066-6
6. Bondi E., Jain R., Aggrawal P., Anand S., Hannaford R., Kapoor A., et al. BIRDSAI: A dataset for detection and tracking in aerial thermal infrared videos. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2020. P. 1736–1745. https://doi.org/10.1109/WACV45572.2020.9093284
7. Dai Y., Wu Y., Zhou F., Barnard K. Asymmetric contextual modulation for infrared small target detection. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2021. P. 949–958. https://doi.org/10.1109/WACV48630.2021.00099
8. Li B., Xiao C., Wang L., Wang Y., Lin Z., Li M., et al. Dense Nested Attention Network for Infrared Small Target Detection. *IEEE Transactions on Image Processing*. 2023. **32**. P. 1745–1758. https://doi.org/10.1109/TIP.2022.3199107
9. Zhang M., Zhang R., Yang Y., Bai H., Zhang J., Guo J. ISNet: Shape Matters for Infrared Small Target Detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022. P. 877–886. https://doi.org/10.1109/CVPR52688.2022.00095
10. Zhang L., Lin W., Shen Z., Zhang D., Xu B., Wang K., et al. Infrared Dim and Small Target Sequence Dataset Generation Method Based on Generative Adversarial Networks. *Electronics*. 2023. **12** (17). P. 3625. https://doi.org/10.3390/electronics12173625
11. Mueller M., Smith N., Ghanem B. A Benchmark and Simulator for UAV Tracking. In: Leibe B., Matas J., Sebe N., Welling M., editors. *Computer Vision – ECCV 2016*. *Lecture Notes in Computer Science*. Vol. **9905**. Cham: Springer; 2016. P. 445–461. https://doi.org/10.1007/978-3-319-46448-0_27
12. Zhang P., Zhao J., Wang D., Lu H., Ruan X. Visible-Thermal UAV Tracking: A Large-Scale Benchmark and New Baseline. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022. P. 8876–8885. https://doi.org/10.1109/CVPR52688.2022.00868
13. Zhu P., Wen L., Du D., Bian X., Fan H., Hu Q., Ling H. Detection and Tracking Meet Drones Challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2022. **44** (11). P. 7380–7399. https://doi.org/10.1109/TPAMI.2021.3119563
14. Svanström F., Alonso-Fernandez F., Englund C. A dataset for multi-sensor drone detection. *Data in Brief*. 2021. **39**. P. 107521. https://doi.org/10.1016/j.dib.2021.107521
15. Kou R., Wang C., Peng Z., Zhao Y., Chen Y., Han J., et al. Infrared small target segmentation networks: A survey. *Pattern Recognition*. 2023. **143**. P. 109788. https://doi.org/10.1016/j.patcog.2023.109788
16. Cheng Y., Lai X., Xia Y., Zhou J. Infrared Dim Small Target Detection Networks: A Review. *Sensors*. 2024. **24** (12). P. 3885. https://doi.org/10.3390/s24123885

17. Xie B., Zhang C., Wang F., Liu P., Lu F., Chen Z., Hu W. CST Anti-UAV: A Thermal Infrared Benchmark for Tiny UAV Tracking in Complex Scenes. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. 2025. P. 6157–6166. arXiv:2507.23473. https://doi.org/10.48550/arXiv.2507.23473
18. Wang Z., Hu Y., Yang J., Zhou G., Liu F., Liu Y. A *Contrastive*-Augmented Memory Network for Anti-UAV Tracking in TIR Videos. *Remote Sensing*. 2024. **16** (24). P. 4775. https://doi.org/10.3390/rs16244775
19. Liu S., Xu T., Zhu X-F., Wu X-J., Kittler J. Learning adaptive detection and tracking collaborations with augmented UAV synthesis for accurate anti-UAV system. *Expert Systems with Applications*. 2025. **282**. P. 127679. https://doi.org/10.1016/j.eswa.2025.127679
20. Dong Y., Wu F., Zhang S., Chen G., Hu Y., Yano M., et al. Securing the Skies: A Comprehensive Survey on Anti-UAV Methods, Benchmarking, and Future Directions. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2025. P. 6725–6739. https://doi.org/10.1109/CVPRW67362.2025.00663
21. Skybytskyi N. Two-Level Algorithm for the UAV Routing Problem with Moving Targets. *Cybernetics and Computer Technologies*. 2025. **4**. P. 29–36. https://doi.org/10.34229/2707-451X.25.4.3

Received/Одержано 06.12.2025 Accepted/Прийнято 26.05.2026 Published/Надруковано 01.06.2026