

Evaluating Trustworthiness of LLMs for Contrail Detection via Ground-Based Cameras

Aizierjiang Aiersilan
The George Washington University
alexandera@gwu.edu

Objective & Project Integration

In this project, taking 2 frontier LLMs (Gemini 3.1 Pro and Claude Opus 4.6) as examples, I explored the trustworthiness of LLMs when deployed as research helpers in a real scientific workflow. Specifically, I integrated them into the contrail project introduced by Prof. Pless, choosing the contrail identification using ground-based visible cameras as an entry point.

The research domain concerns aviation condensation trails (contrails), which studies estimate to account for roughly 35% of aviation’s total climate forcing [6]. While recent studies has mainly focused the approaches from the satellite imagery, many challenges still exist and ground-based camera systems as is mentioned in the paper that proposes GVCCS dataset [5] and the contrail-to-flight attribution pipeline of Dalmau et al. [3] offer a promising replacement to satellite-only observation, providing high temporal and spatial resolution imagery captured moments after contrail formation.

My workflow proceeded in 3 stages:

1. **Brief Literature Review (Gemini 3.1 Pro):** I prompted Gemini to synthesize key findings from published papers and authoritative web resources on contrail climate impact and ground-camera solutions, requiring verbatim source quotations for every factual claim.
2. **Experimental Design (Gemini 3.1 Pro):** I prompted Gemini to propose a complete pipeline for differentiating real contrails from optical glares in whole-sky imagery, including neural architecture selection and ADS-B data fusion.
3. **Adversarial Cross-Model Audit (Claude Opus 4.6):** I fed Gemini’s experimental design to a second, independent LLM (Claude Opus 4.6) and asked it to identify technical pitfalls, physical impossibilities, and architectural biases.

This workflow was designed to test a core question in trustworthy AI: *Can LLMs be reliable collaborators in scientific research, and what safeguards are necessary to prevent confidently stated but technically flawed outputs from entering the research pipeline?*

AI in Short Literature Review

To evaluate LLM’s accuracy in scientific summarization, I uploaded the the paper [5] that proposes the GVCCS dataset, the contrail-to-flight attribution paper [3] (all in PDF format), and a brief screenshot of the GVCCS dataset structure to Gemini 3.1 Pro, along with three web resource links. I deliberately constrained the model to append a direct, verbatim quote from the source material for every factual claim so that I can audit its output. My full prompt provided to Gemini 3.1 Pro is presented in detail in Appendix B, while the full response from Gemini 3.1 Pro is presented in Appendix C and my manual verification is provided in Appendix A.

AI in Experimental Design

Having established Gemini’s reliability for literature synthesis, I then tested a higher-risk capability by prompting Gemini 3.1 Pro to propose a complete pipeline for differentiating real contrails from optical glares using the whole-sky sequences from the GVCCS dataset. (See Appendix B for full prompt for experimental design and Appendix C for its full response).

At first glance, the pipeline it provided appears technically sophisticated and well-structured, leveraging SOTA architectures (Mask2Former, Swin Transformers), incorporating temporal reasoning, and proposing a multi-modal validation scheme. However, as the cross-model audit in the next section reveals, this confident presentation may conceal several severe technical limitations or even flaws (under the assumption that Claude opus 4.6 is able to pick out such flaws accurately).

Cross-Model Critique

To test the reliability of Gemini’s experimental design, I submitted it to Claude Opus 4.6 which is another independent frontier LLM and asked it to perform an adversarial technical review. (See Appendix B for my full prompt and Appendix C for its full critique).

Claude Opus 4.6 claims in its critique that Gemini’s altitude threshold of 15,000 ft for filtering ADS-B trajectories is **physically inappropriate** and it also states that sun glare in a fisheye lens is not a single, smoothly tracking bright spot. It is a complex, spatially distributed and temporally erratic phenomenon composed of several distinct components. Though they seem generally reasonable and correct, providing critical insights for thorough consideration of the pipeline generated by Gemini 3.1 Pro, it is clearly nitpicking in most aspects (e.g. “GVCCS dataset lacks thermal infrared channels”, “Contrail formation follows the Schmidt-Appleman criterion [8], requiring 26,000–30,000,ft mid-latitude temperatures, so flights at 15,000–26,000,ft inject systematic false positives into validation”).

And when it comes to Neural Architecture Pitfalls, Claude Opus 4.6 claims to have identified 3 architectural biases in Gemini’s proposed pipeline (See Appendix C for full response):

1. **Swin window partitioning dilutes thin-structure features.** Contrails are often 1–5 pixels wide but span hundreds of pixels. The local window attention mechanism of Swin Transformers [7] means contrail pixels occupy a tiny fraction of tokens in each window, diluting their signal.
2. **Insufficient temporal receptive field.** With GVCCS capture intervals of 15–60 seconds, a 5-frame clip spans only 1–5 minutes, during which the sun moves approximately 0.25° – 1.25° in azimuth, often below the noise floor for glare position estimation.
3. **Mask2Former’s query-based matching is biased against contrails.** The dice loss used in bipartite matching during training is highly sensitive to small lateral shifts of thin masks, biasing the network toward detecting thicker, more compact objects [2].

Claude also critiqued the choice of Hausdorff distance as outlier-sensitive and inappropriate for fragmented structures, the unquantified uncertainty in ERA5 advection modeling (spatial resolution of ~ 31 km, temporal resolution of 1 hour), and the information loss from fisheye-to-Cartesian rectification at horizon regions where contrails are most scientifically important. Claude also noted missing elements including the absence of ablation experiments, no discussion of extreme class imbalance (contrail pixels often constitute $<0.1\%$ of total image area), and no explicit glare augmentation strategy.

It seems that my prompt to critique unintentionally opened the debating strategy of Claude, since it’s apparently finding some clues and facts that may support its claim to oppose Gemini’s response as much as it could. But I think such kinds of oppositions and critiques are still valuable

for human researchers by providing quick exploration of the limitations of the approach. (while the fact that such leaning of AI-critique may or may not lead to cognitive offloading which may gradually reduces the independent critical thinking abilities of the researchers is another problem).

Broader Implications for Trustworthy AI

I noticed a gap between *fluency* and *correctness*. When constrained to provided sources, Gemini 3.1 Pro achieves 100% grounding, confirming its value as a powerful literature review assistant [1]. However, in its 17th claim, it mirrors satellite imagery cost-efficiency directly to ground-camera solutions (See Appendix A for more details). If I were the writer, I would clarify that deploying a network of 10,000 ground cameras (amortized at ~\$150 million annually) adds only about \$1.50 per ton of CO₂e avoided. Even when combined with extra fuel costs, the total (\$10–\$30/ton) remains a massive bargain compared to Sustainable Aviation Fuel (>\$200/ton) or Direct Air Capture (>\$600/ton). I would further add that leveraging existing public webcams could drive this cost down even more drastically, presenting reviewers with an undeniably compelling economic proposition. While I agree with the model’s summary, I feel it lacks persuasive punch. It should explicitly highlight satellite weaknesses and ground-camera advantages, adopting a narrower scope like the introduction of [3]. So I still think reading the original papers introduction sections is more efficient and clear for grasping the full picture.

When the task shifted from summarization to *generation of novel scientific methodology*, the model produced outputs containing some limitations that were marked as “fundamental physical errors,” “architecturally mismatched components,” and “a critically flawed evaluation framework” by another frontier LLM. While such a cross-model audit demonstrates one promising mitigation strategy of adversarial multi-model oversight by treating one LLM’s output as input for critique by an independent model, a form of automated peer review that can catch errors that neither model would catch in its own output was thus created, paralleling with ensemble methods in machine learning [4] and with the broader principle that diversity of perspective improves error detection.

While it’s often good for researchers to be alerted about potential limitations in their proposals, LLM-designed experimental pipelines can sometimes render solutions unexpectedly more complicated as is the case in this study. Adopters must therefore deconstruct [9] these designs to uncover and apply only parts that really matters, or try simpler, more effective approaches to challenging problems.

Multi-model review alone proves insufficient. Although Claude’s critique was technically impressive, it was itself an LLM-generated output that could contain subtle errors or overlook nuances a domain expert would immediately notice and its critique against the Gemini’s response is apparently nit picking and sometimes even using some sources that are not “trustworthy” at all (e.g. a paper from a low IF OpenAccess journal, a low-quality blog post that have millions of viewers, etc.) to accomplish the opposition and critique. I conclude from these tests that an appropriate framework may well be an adversarial human-in-the-loop oversight augmented by multi-model cross-checking and deconstruction: LLMs function as brainstorming partners whose outputs undergo automated cross-model critique before receiving final human expert validation. And the trustworthy deployment of AI in science requires not less human judgment but *differently structured* human judgment, with less time spent on literature searches and more time devoted to the critical evaluation of AI-generated experimental designs.

Acknowledgments

I thank Prof. Robert Pless for initiating this assignment in Trustworthy AI course taught at the George Washington University for Spring 2026.

References

- [1] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.
- [2] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022.
- [3] Ramon Dalmau, Gabriel Jarry, and Philippe Very. Contrail-to-flight attribution using ground visible cameras and flight surveillance data. *arXiv preprint arXiv:2510.16891*, 2025.
- [4] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.
- [5] Gabriel Jarry, Ramon Dalmau, Philippe Very, Franck Ballerini, and Stefania-Denisa Bocu. Gvccs: a dataset for contrail identification and tracking on visible whole sky camera sequences. *Earth System Science Data*, 18(2):1037–1059, 2026.
- [6] David S Lee, David W Fahey, Agnieszka Skowron, Myles R Allen, Ulrike Burkhardt, Qi Chen, Sarah J Doherty, Sarah Freeman, Piers M Forster, Jan Fuglestedt, et al. The contribution of global aviation to anthropogenic climate forcing for 2000 to 2018. *Atmospheric environment*, 244:117834, 2021.
- [7] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.
- [8] Ulrich Schumann. On conditions for contrail formation from aircraft exhausts. *Meteorologische Zeitschrift*, 5:4–23, 1996.
- [9] Kilian Weinberger. The importance of deconstruction. Presentation at the ML Retrospectives, Surveys & Meta-Analyses (ML-RSA) Workshop, NeurIPS, December 2020. URL <https://slideslive.com/38938218/the-importance-of-deconstruction>. SlidesLive Video.

Accuracy Audit of Literature Review

Manual Verification of Quoted Claims

To ensure the accuracy of Gemini’s literature review, each quoted passage was manually checked against the original source documents. The results of my verification are summarized below.

Source Verification Results (17 Claims)

1. “Aviation contributes to global climate change...water vapour and aerosols.” — **Verified** in GVCCS Dataset paper, Introduction.
2. “Only slightly over half of aviation warming comes from carbon dioxide” — **Verified** at <https://sites.research.google/gr/contrails/>, image caption.
3. “Condensation trails, or contrails, are created by an aircraft’s exhaust...” — **Verified** at Northrop Grumman article, second sentence.
4. “A recent IPCC report noted that clouds created by contrails account for roughly 35%...” — **Verified** at Google Research contrails page, second paragraph.
5. “These ice clouds can alter Earth’s radiative balance...” — **Verified** in GVCCS Dataset paper, Abstract.
6. “When it persists, a contrail acts similar to an ice cloud...” — **Verified** in MIT News article, “Imaging backbone” section.
7. “At night, the cloud-like threads have only a trapping, warming effect.” — **Verified** in MIT News article, “Imaging backbone” section.
8. “Recent studies suggest that the climate forcing...” — **Verified** in GVCCS Dataset paper, Introduction.
9. “predicting persistent contrails is challenging...” — **Verified** at Northrop Grumman article, “Cloudy with a Chance of Meatballs” section.
10. “Geostationary satellite images are the workhorse...” — **Verified** in MIT News article, “Imaging backbone” section.
11. “Images from geostationary satellites alone aren’t enough...” — **Verified** in MIT News article, subtitle.
12. “We found 80 percent of the contrails...” — **Verified** in MIT News article, “Catching the trail” section.
13. “By the time contrails are visible at satellite resolution...” — **Verified** in Contrail-to-Flight Attribution paper, Introduction.
14. “Ground-based cameras offer a promising alternative...” — **Verified** in Contrail-to-Flight Attribution paper, Introduction.
15. “This high temporal and spatial resolution reduces attribution ambiguity...” — **Verified** in Contrail-to-Flight Attribution paper, “State of the Art” section.
16. “GVCCS contains video sequences recorded by a ground-based camera in Brétigny-sur-Orge...” — **Verified** in Contrail-to-Flight Attribution paper, Introduction.
17. “reduced contrails by 54%...”, “flights that avoided contrails burned 2% more fuel”, “contrail avoidance costs could be in the range of \$5–25/ton CO₂e” — **Verified** at Google Research contrails page, “Testing with American Airlines” section.

My Prompts

Prompt

[Attached files: paper - GVCCS Dataset.pdf, paper - Contrail-to-Flight Attribution Using Ground Visible Cameras and Flight Surveillance Data.pdf, screenshot_of_GVCCS_Dataset]

I'm researcher in computer vision and now researching contrail identification using Ground Visible Cameras. Please act as a senior climate tech researcher. Review the provided GVCCS dataset paper, the paper – Contrail-to-Flight Attribution Using Ground Visible Cameras and Flight Surveillance Data and the link provided below that describes MIT study on this topic. Write a 500-word synthesis of the environmental impact of contrails and how ground cameras solve this. Constraint: You must append a direct, verbatim quote from the text in parentheses for every factual claim or statistic you make. You must be highly accurate and precise, logically consistent and smooth, solid and correct. You should make full use of all the resources that I provided and make sure you reach your whole potential to be accurate and insightful.

<https://sites.research.google/gr/contrails/>

<https://www.northropgrumman.com/life-at-northrop-grumman/clearing-the-skies>

<https://news.mit.edu/2025/study-observing-skies-will-help-planes-reduce-climate-warming-contrails-1222>

[Note that for the GVCCS dataset, I've already downloaded them to my local disk and attached screenshot shows the brief structure of the images in the dataset, for your reference if necessary.]

Prompt to Gemini 3.1 Pro (Experimental Design)

Propose a complete computer vision experimental design to differentiate real contrails from optical glares using the fisheye whole-sky sequences from the GVCCS dataset. Outline the specific architecture (e.g., Vision Transformers, 3D CNNs) needed to capture the temporal movement of clouds versus the sun-tracking movement of glares. Finally, explain how to fuse ADS-B flight surveillance data into this pipeline to validate the visual predictions.

Prompt to Claude Opus 4.6

Critique this proposed computer vision experimental design. Identify any technical pitfalls, physical impossibilities regarding optics/glaires, or biases in the suggested neural network architecture.

[Gemini 3.1 Pro's full experimental design was provided as context.]

Gemini 3.1 Pro — Literature Synthesis

As a senior climate tech researcher, I recognize that mitigating aviation's environmental footprint requires addressing its most misunderstood byproduct: contrails. Aviation's climate impact extends far beyond standard greenhouse gases ("Aviation contributes to global climate change not only through carbon dioxide (CO₂) emissions but also through a variety of non- CO₂ effects, including nitrogen oxides (NO_x), water vapour and aerosols."). In fact, the majority of the industry's warming effect is not from carbon ("Only slightly over half of aviation warming comes from carbon dioxide"). The visible lines planes leave behind ("Condensation trails, or contrails, are created by an aircraft's exhaust and can trap energy in the Earth's atmosphere.") are highly consequential. A major scientific consensus highlights their severe planetary impact ("A recent IPCC report noted that clouds created by contrails account for roughly 35% of aviation's global warming impact."). The warming mechanism of these trails is complex but powerful. These artificial clouds disrupt the Earth's natural temperature regulation ("These ice clouds can alter Earth's radiative balance, potentially rivaling the warming effect of aviation CO₂"). They possess dual, conflicting properties during the daytime ("When it persists, a contrail acts similar to an ice cloud and, as such, can have two competing effects: one in which the contrail is a sort of heat shield, reflecting some incoming radiation from the sun. On the other hand, a contrail can also act as a blanket, absorbing and reflecting back some of the heat from the surface."). However, post-sunset, the reflective cooling effect vanishes ("At night, the cloud-like threads have only a trapping, warming effect."). As a result, the net outcome is a rapidly warming planet ("Recent studies suggest that the climate forcing due to contrail cirrus clouds is of the same order of magnitude as aviation CO₂ emissions").

Solving this requires contrail-to-flight attribution, but tracking these formations is notoriously difficult ("predicting persistent contrails is challenging because they form only under specific atmospheric conditions"). Historically, scientists relied heavily on high-altitude observations ("Geostationary satellite images are the workhorse of observations for detecting contrails"). However, these legacy methods are fundamentally inadequate on their own ("Images from geostationary satellites alone aren't enough to help planes avoid contrail-prone regions"). In fact, massive amounts of data are missed by standard satellite systems ("We found 80 percent of the contrails we could see with LEO satellites, we couldn't see with GEO imagers"). Furthermore, space-based imagery suffers from severe lag and spatial distortion ("By the time contrails are visible at satellite resolution, they have often drifted, deformed, or merged with other clouds, making alignment with flight trajectories highly uncertain").

This is where ground-based visible cameras provide a revolutionary solution. Ground sensors fix the temporal and spatial resolution problems of satellites ("Ground-based cameras offer a promising alternative. They capture contrails just moments after formation, at high spatial and temporal resolution, when they are still thin, linear, and visually distinct."). By capturing the exhaust plumes immediately, researchers can accurately match the visual data to the specific aircraft ("This high temporal and spatial resolution reduces attribution ambiguity and allows direct association with flight trajectory data"). Dedicated dataset projects are actively pioneering this approach ("GVCCS contains video sequences recorded by a ground-based camera in Brétigny-sur-Orge, France, with detailed annotations for each contrail").

With accurate ground truth, AI models can successfully guide pilots to bypass humid regions. Real-world testing proves this works ("reduced contrails by 54% compared to when pilots didn't use our predictions."). While altering flight paths requires slightly more

energy (“flights that avoided contrails burned 2% more fuel”), the financial viability of this strategy is unmatched (“contrail avoidance costs could be in the range of \$5–25/ton CO_{2e}”). By deploying ground cameras to validate predictive models, the aviation industry can finally eliminate one of its largest climate impacts.

Gemini 3.1 Pro — Proposed CV Pipeline

Designing an experimental pipeline to isolate real contrails from optical glares in fisheye whole-sky imagery is a classic, high-impact computer vision challenge. Sun glare and low-altitude clouds frequently interrupt contrail continuity and introduce visually ambiguous false positives.

To effectively tackle this for a top-tier venue submission, the design must decouple the visual kinematic profile—wind-driven advection for clouds versus deterministic solar tracking for glares—and then ground the visual predictions in deterministic flight data.

Here is a proposed experimental architecture using the GVCCS dataset.

Phase 1: Spatiotemporal Visual Architecture (The Vision Baseline)

A purely 2D instance segmentation model will struggle to separate static or slowly tracking glares from drifting contrails. You need a Video Instance Segmentation (VIS) framework that treats time as a core feature dimension.

- (1) Architecture Choice: Implement a video-variant of Mask2Former with a 3D Swin-Large (Swin-L) backbone. This architecture is highly effective because it replaces global attention with masked attention, focusing the transformer’s learnable queries specifically on the boundaries of thin, high-aspect-ratio structures.
- (2) Temporal Clip Processing: Train the network on overlapping short video clips (e.g., N=3 to 5 frames) utilizing a sliding window approach with a stride of 1. By processing multi-frame volumes, the model captures the delta in pixel displacement.
- (3) Feature Discrimination: Sun glares move predictably with the solar azimuth, whereas contrails advect according to high-altitude wind fields. The 3D Swin-L backbone will implicitly learn these differing motion vectors, allowing the queries to suppress glare artifacts as background “stuff” rather than instance-level “things.”

Phase 2: Post-Processing & Instance Linking

Because the video-based Mask2Former processes short clips independently, you must maintain instance identities across the full video sequence to map the complete trajectory.

- (1) Bipartite Matching: Extract the predicted masks and compute the negative temporal Intersection over Union (IoU) across overlapping frames.
- (2) Hungarian Algorithm: Use the Hungarian algorithm on this cost matrix to optimally link disjointed contrail fragments across the sequence. Glares that intermittently flash or shift will fail to establish a temporally consistent track and can be filtered out.

Phase 3: ADS-B Fusion & Kinematic Validation (The Ground Truth Filter)

Even with robust spatiotemporal visual modeling, some atmospheric phenomena or persistent glares might trick the network. Fusing ADS-B trajectory data serves as the ultimate deterministic filter.

- (1) Generate Theoretical Contrails: Extract the ADS-B flight trajectories filtered for cruising altitudes (e.g., above 15,000 ft). Use a dry advection model and ERA5 meteorological wind data to simulate how the exhaust plume drifts from its emission point.
- (2) Geometric Projection: Project these simulated 3D advected polygons into the 2D Cartesian pixel coordinates of the GVCCS geometrically projected images.

(3) Spatial Matching: Compute the agreement between your visual Mask2Former predictions (A) and the theoretical advected masks (B) using a directed Hausdorff distance: $d(A \rightarrow B) = \max_{a \in A} \min_{b \in B} \|a - b\|$

(4) The Glare Rejection: Apply a maximum acceptable separation threshold, τ_d . If a tracked visual object (like a persistent glare) has a high Hausdorff distance or lacks any matching ADS-B theoretical trajectory within τ_d , the assignment probability drops to zero, and the visual detection is classified as a false positive and discarded.

Claude Opus 4.6 — Circular Validation Critique

Comprehensive Critique of the Proposed Experimental Design

This is a well-structured proposal that demonstrates genuine understanding of the problem domain. However, it contains several significant technical pitfalls, physical misunderstandings, and architectural biases that would undermine both the scientific validity and the practical performance of the system. I will address these systematically.

Optical and Physical Misunderstandings

The Glare Motion Model Is Fundamentally Wrong

The proposal’s core discriminative thesis—that “sun glares move predictably with the solar azimuth, whereas contrails advect according to high-altitude wind fields”—reveals a flawed physical model of how glare actually manifests in whole-sky imagery.

The problem: Sun glare in a fisheye lens is not a single, smoothly tracking bright spot. It is a complex, **spatially distributed and temporally erratic** phenomenon composed of several distinct components:

- **Veiling glare (sensor-level):** A diffuse luminance wash across large portions of the sensor caused by internal scattering within the lens barrel and between lens elements. This does not “move” in a trackable sense; it modulates the entire image’s contrast non-uniformly and changes character as the sun crosses different regions of the fisheye’s highly curved element stack.
- **Lens flare artifacts:** These are discrete bright spots and streaks caused by inter-element reflections. Critically, their positions are determined by the geometry of the specific lens assembly. They move **along a line through the image center and the sun’s position**, but their number, intensity, and spacing change discontinuously as the sun’s angle relative to the optical axis changes. They do not follow a smooth, easily predicted trajectory.
- **Specular bloom and diffraction spikes:** Caused by the aperture blades and any protective housing or dome over the camera. These rotate and reshape based on subtle mechanical and atmospheric factors.
- **Atmospheric forward scattering (aureole):** The bright halo around the sun caused by aerosol and ice crystal scattering. This is atmospheric, not optical, and its size and intensity depend on aerosol optical depth, which varies on timescales of minutes to hours.

The consequence: The proposal assumes the network will learn a single, clean “solar

tracking” motion vector for glare. In reality, different glare components exhibit different spatial behaviors simultaneously—some static relative to the sun, some jumping discontinuously, some diffuse and non-localizable. A 3D convolutional backbone processing 3–5 frame clips will not cleanly decompose this mixture into a single suppressible motion pattern. Some glare components, particularly veiling glare that reduces local contrast, will instead cause the network to **miss contrails that pass through the glare region** rather than hallucinate false contrails.

Contrail Appearance Is Not Purely Kinematic

The proposal implicitly treats contrails as having a consistent visual signature that merely needs to be tracked. In reality, contrails in whole-sky imagery exhibit extreme appearance variability:

- **Optical thickness varies along the contrail’s length** as ice crystal growth and sublimation depend on local supersaturation. A single contrail can be bright and opaque in one segment and nearly invisible in another.
- **Contrails frequently overlap with or are embedded within cirrus cloud layers**, making them indistinguishable from natural cirrus in single-band visible imagery. The GVCCS dataset uses visible-band cameras, which lack the thermal infrared channels (particularly the 10.8–12.0 μm brightness temperature difference) that are the standard tool for contrail–cirrus discrimination in satellite remote sensing.
- **Aging contrails spread and lose their linear morphology** within 30–120 minutes, becoming diffuse contrail-cirrus that no instance segmentation model trained on linear features will detect.

This means the proposal’s framing of the problem as “contrail vs. glare” is an incomplete taxonomy. The harder and more scientifically important discrimination is **contrail vs. natural cirrus vs. contrail-cirrus**, which this pipeline does not address at all.

The 15,000 ft Cruising Altitude Filter Is Physically Inappropriate

The proposal filters ADS-B trajectories for aircraft “above 15,000 ft” to isolate potential contrail-producing flights. This threshold is far too low and introduces systematic false positives in the validation phase.

Physical basis: Contrail formation is governed by the Schmidt–Appleman criterion, which requires that the ambient atmosphere be cold and moist enough for the mixture of hot, moist exhaust and ambient air to momentarily reach water saturation. In practice, this almost never occurs below approximately **8,000–9,000 metres (roughly 26,000–30,000 ft)** in mid-latitudes, and the threshold is even higher in tropical regions. Aircraft at 15,000 ft are typically in conditions far too warm for contrail formation.

The consequence: By including flights between 15,000 and $\approx 26,000$ ft, Phase 3 will generate a large number of “theoretical contrails” from flights that could not possibly have produced contrails. These phantom theoretical masks will either (a) fail to match any visual detection, adding noise to the validation statistics, or (b) worse, coincidentally overlap with natural cirrus or glare artifacts, causing the system to **validate false positives** as real contrails.

A corrected design should either apply the Schmidt–Appleman criterion directly using ERA5 temperature and humidity profiles at flight level, or at minimum raise the altitude filter to 28,000 ft as a conservative proxy.

Neural Network Architecture Pitfalls and Biases

The 3D Swin-L Backbone Is Poorly Matched to the Data Geometry

The choice of a 3D Swin-Large backbone is presented as a natural extension for temporal modelling, but it introduces several problems specific to this domain.

Shifted window bias against thin structures: The Swin Transformer’s core mechanism partitions the feature map into non-overlapping local windows and computes self-attention within each window, then shifts the windows to enable cross-window communication. Contrails in whole-sky imagery are **extremely thin (often 1–5 pixels wide) and extremely long (potentially spanning hundreds of pixels across the image)**. A contrail that runs diagonally across the image will intersect many windows but occupy only a tiny fraction of the tokens within each window. The local attention within each window will be dominated by background sky or cloud tokens, and the contrail signal will be diluted.

The proposal claims that masked attention in Mask2Former solves this by “focusing queries on the boundaries of thin, high-aspect-ratio structures.” This is misleading. Masked attention in Mask2Former means that each query attends only to the spatial region predicted by its mask from the previous decoder layer. This is beneficial for separating instances but does **not** inherently help with detecting thin structures in the first place—the initial mask predictions from the first decoder layer still depend on the backbone features, which suffer from the window partitioning problem described above.

A more appropriate backbone choice would be a hybrid architecture that includes explicit long-range connectivity for thin structures, such as incorporating deformable attention layers (as in Deformable DETR) that can attend to sparse, distant locations along a linear feature, or using an explicit line-detection prior.

The 3D Convolution Temporal Receptive Field Is Insufficient

Processing clips of $N = 3$ to 5 frames with a stride of 1 frame provides an extremely limited temporal receptive field. The proposal claims this is sufficient for the network to “implicitly learn differing motion vectors” between glare and contrails.

Quantitative analysis: In typical whole-sky camera setups (including those in GVCCS), images are captured at intervals of 15–60 seconds. A 5-frame clip therefore spans 1–5 minutes. During this time:

- The sun moves approximately 0.25° – 1.25° in azimuth (at mid-latitudes), which in a fisheye image corresponds to only a few pixels of glare displacement. This is **below the noise floor** of glare position estimation given the spatially diffuse nature of most glare artifacts.
- A contrail advecting in a 30 m/s upper-level wind moves approximately 450–9,000 metres horizontally, which depending on the camera’s field of view and the contrail’s altitude, may correspond to anywhere from a few pixels to tens of pixels.

The motion differential between glare and contrail within a 3–5 frame window is therefore **marginal and often ambiguous**, particularly for young contrails near the sun (which is exactly where glare contamination is worst).

Mask2Former’s Query-Based Architecture Has a Systematic Bias Against Contrails

Mask2Former uses a fixed number of learnable queries (typically 100–200), each of which is trained to detect one instance. This creates two problems:

- **Contrail multiplicity:** A single whole-sky image can contain dozens of contrails simultaneously, some overlapping or nearly parallel. If the number of queries is set too low, the model will systematically under-detect in high-traffic airspace. If set too high, training becomes unstable due to the large number of unmatched queries.
- **Aspect ratio bias in the Hungarian matching during training:** During training, Mask2Former uses bipartite matching to assign ground truth masks to queries. The matching cost typically includes a dice loss or mask IoU term. For extremely thin, elongated masks (contrails), the dice score is highly sensitive to even small spatial misalignments—a 2-pixel lateral shift of a 3-pixel-wide contrail mask drops the dice score dramatically. This makes the training matching unstable and biases the network toward detecting thicker, more compact objects (like cloud patches or glare blobs) where the matching is more forgiving.

Methodological and Evaluation Biases

The Hausdorff Distance Is a Poor Choice for Spatial Matching

The proposal uses the directed Hausdorff distance $d(A \rightarrow B)$ to match visual predictions to theoretical ADS-B contrail masks. This metric is **maximally sensitive to outliers** by definition—it reports the maximum over all points in A of the minimum distance to B . **The problem:** Contrails in imagery are often fragmented due to gaps in supersaturation, cloud occlusion, or glare interference. A single outlier pixel in the predicted mask (e.g. a small segmentation artifact 50 pixels away from the true contrail) will dominate the Hausdorff distance and potentially cause a valid detection to be rejected. Conversely, a glare streak that happens to run roughly parallel to a flight path could have a low Hausdorff distance despite being physically unrelated.

A better choice would be the **average Hausdorff distance** or, more appropriately for this elongated-structure matching problem, a **Fréchet distance** or **curve-based chamfer distance** that respects the sequential geometry of the contrail and is robust to fragmentation.

Circular Validation Creates a Confirmation Bias

The most serious methodological flaw in the entire pipeline is that **Phase 3 is presented as both a validation mechanism and a filtering step, but it cannot serve both roles simultaneously.**

Here is the logical structure:

1. The visual model (Phases 1–2) produces candidate contrail detections.
2. Phase 3 filters these detections by requiring agreement with ADS-B-derived theoretical contrails.
3. The proposal implies that the remaining detections (those that pass the filter) constitute the system’s output and that the filtering step validates the visual model’s accuracy.

The circularity: If you discard all visual detections that lack ADS-B support, you can never measure the visual model’s **false negative rate** for contrails that exist but lack ADS-B coverage (ADS-B coverage is not universal—military aircraft, some general aviation, and aircraft over oceanic regions without ground-based receivers may be absent). You also cannot measure the visual model’s ability to detect **contrail-cirrus** (aged, spread contrails) because these will have drifted far from any ADS-B trajectory and will be rejected by the Hausdorff threshold.

More fundamentally, the system’s **precision** becomes artificially inflated because you are

defining “true positive” as “agrees with ADS-B” rather than using independent ground truth. This is not validation; it is **filtering by assumption**.

A rigorous experimental design must include an **independent evaluation set** with manually annotated ground truth masks (including difficult cases like contrails in glare, contrails overlapping cirrus, and glare artifacts that resemble contrails) and must report precision, recall, and F1 against this independent annotation, not against the ADS-B fusion output.

The Advection Model Introduces Unquantified Uncertainty

The proposal uses ERA5 reanalysis wind fields to advect theoretical contrail plumes. ERA5 has a spatial resolution of ~ 31 km and a temporal resolution of 1 hour. Upper-tropospheric wind fields can vary significantly on scales smaller than this, particularly near jet streams where wind shear is strong—which is also exactly where contrails are most common and most persistent.

The consequence: The theoretical advected contrail positions will have spatial errors of potentially **several kilometres**, which after projection into the fisheye image could correspond to **tens of pixels** of positional uncertainty. The Hausdorff distance threshold τ_d must be set large enough to accommodate this uncertainty, but a large threshold also admits more false matches with glare or cirrus. The proposal does not discuss how to set τ_d or how to propagate the meteorological uncertainty into the matching confidence.

The Fisheye-to-Cartesian Projection Step Discards Information

The proposal mentions projecting into “2D Cartesian pixel coordinates of the GVCCS geometrically projected images.” This implies the fisheye images are being rectified (unwarped) to a Cartesian projection before processing.

The problem: Fisheye rectification introduces **spatially varying distortion and resolution loss**, particularly near the horizon where the fisheye mapping compresses many degrees of sky into few pixels. Contrails near the horizon (which are the most distant and therefore the longest-lived and most scientifically interesting for climate studies) will be severely degraded by this rectification. Furthermore, the rectification changes the effective point spread function across the image, meaning the network must implicitly learn a spatially varying appearance model.

A more principled approach would be to either (a) work directly in the native fisheye projection and project the ADS-B trajectories into fisheye coordinates (which is straightforward given a calibrated lens model), or (b) use a **polar or equidistant projection** that better preserves the angular uniformity of the sky hemisphere.

Missing Elements for a Top-Tier Submission

No Ablation Structure

The proposal describes a single monolithic pipeline but does not specify ablation experiments. A top venue will require demonstrating the marginal contribution of each component: the temporal backbone vs. a 2D baseline, the Hungarian linking vs. simple IoU tracking, the ADS-B fusion vs. vision-only, etc.

No Discussion of Class Imbalance

In whole-sky imagery, contrail pixels constitute a tiny fraction of the total image area (often $< 0.1\%$). The proposal does not discuss how to handle this extreme class imbalance

during training—no mention of focal loss, hard example mining, or oversampling strategies.

No Glare Augmentation or Explicit Glare Modelling

If the goal is to make the network robust to glare, the proposal should include explicit glare simulation as a data augmentation strategy during training—synthetic glare overlaid on contrail annotations to force the network to learn to see through it. The current design relies entirely on the network implicitly learning glare robustness from whatever glare examples happen to exist in the training set, which is unlikely to cover the full diversity of glare conditions.

Summary of Critical Issues

| Category | Issue | Severity |
|--------------|---|-------------|
| Physics | Glare is not a single trackable object; it is a multi-component optical phenomenon | High |
| Physics | 15,000 ft altitude filter is far below the contrail formation threshold | High |
| Architecture | Swin window partitioning dilutes thin-structure features | Medium-High |
| Architecture | 3–5 frame temporal window is insufficient to discriminate motion regimes | Medium |
| Architecture | Query-based matching is biased against thin, elongated masks | Medium |
| Methodology | Hausdorff distance is outlier-sensitive and inappropriate for fragmented structures | Medium |
| Methodology | ADS-B fusion creates circular validation; no independent ground truth | Critical |
| Methodology | ERA5 advection uncertainty is unquantified and potentially large | Medium |
| Methodology | Fisheye rectification degrades horizon regions where contrails are most important | Medium |

The proposal demonstrates a solid conceptual framework, but its execution plan contains physical misunderstandings about glare behaviour, an architecture that is subtly mismatched to the geometry of the target features, and a validation methodology that cannot distinguish genuine performance from confirmation bias. Addressing these issues would require rethinking the glare model, redesigning the backbone for thin-structure sensitivity, and most importantly, establishing an independent annotated evaluation protocol that is decoupled from the ADS-B fusion pipeline.