https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

Contrails cannot exist without flights: physics-informed contrail detection, tracking, and attribution in ground-based camera video

Ramon Dalmau[1], Gabriel Jarry[1], and Philippe Very[1]
[1]EUROCONTROL Innovation Hub, Brétigny-sur-Orge, France
Correspondence: Ramon Dalmau (ramon.dalmau-codina@eurocontrol.int)

Abstract. A contrail cannot exist without a flight. We exploit this constraint by inverting the usual detect-then-attribute pipeline: instead of finding contrails in an image and asking which flight made them, we start from each flight and ask whether its contrail is visible. Aircraft positions from surveillance data and reanalysis wind fields feed a contrail lifecycle model that predicts where each flight's contrail should appear in a ground-based camera image, not at the aircraft's position but where the wind has carried the plume. These predictions become spatial prompts that guide a video segmentation model to either outline the contrail or reject the prompt when nothing is visible. Because each prompt belongs to exactly one flight, the output mask carries the flight's identity by construction, and tracking across frames requires no re-identification. We compare prompt encodings of increasing physical richness: binary presence, age-weighted freshness, and suppression of competing flights. On the Ground Visible Camera Contrail Sequences dataset, the richest design improves mean average precision, averaged over intersection-over-union matching thresholds from 0.25 to 0.75, by 25% over the binary baseline, reaches 96.2% attribution precision, rejects empty prompts with an area under the receiver operating characteristic curve of 0.97, and maintains detection across roughly three-quarters of a contrail's visible lifetime. A video-level cluster bootstrap confirms the attribution and segmentation-quality gains, and is equally clear that the richer prompts do *not* find more contrails: detection coverage is statistically indistinguishable from the binary baseline. What richer prompts buy is delineation and identity, not discovery.

1 Introduction

Commercial aircraft at cruise altitude produce condensation trails when hot exhaust mixes with the cold, dry air near the tropopause. Whether a contrail persists depends on atmospheric humidity (Sect. 2.1), and the ones that do persist spread into cirrus sheets that trap outgoing infrared radiation. Recent assessments put aviation contrail forcing on par with cumulative CO_2 emissions from aviation (Lee et al., 2021; Teoh et al., 2023).

Mitigation is possible because the impact is concentrated. Only 2 to 10% of flights produce most of the contrail climate impact (Teoh et al., 2020), since formation and persistence depend on atmospheric conditions that vary sharply with altitude and location. Adjusting the altitude of those specific flights by 500 to 1,000 feet could reduce aviation's warming contribution at negligible fuel cost, and operational trials have now shown that airlines can execute such diversions within normal dispatch procedures (Sonabend-W et al., 2024; Sankar et al., 2026).

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

Those same trials show why observation, not just prediction, is the binding constraint. Both of them measured their outcome with satellite imagery and an automated flight-contrail attribution algorithm rather than with the forecast that triggered the diversion, because the forecast is what is under test. The most recent randomized trial reports a 62% lower contrail formation rate among the flights that actually flew the avoidance plan, but only an 11.6% reduction under intent-to-treat over 1,232 eligible flights (Sankar et al., 2026). A gap of that size between planned and realized effect is exactly what independent observation is for: it separates a forecast that was wrong from a manoeuvre that was never flown. Meteorological inputs at cruise altitude, especially humidity, carry enough uncertainty that formation and persistence criteria cannot be taken on trust at the level of an individual flight. Monitoring can draw on satellite imagery, which covers wide areas but at coarser resolution and with a time lag, or on ground-based cameras, which resolve contrails shortly after formation over a small patch of sky. Either way it requires three tightly linked capabilities: detecting contrails in the imagery, tracking them across frames as they evolve, and attributing each contrail to the flight that produced it.

The coupling among these three tasks is the central difficulty. Detection must distinguish contrail pixels from natural cirrus, sun glints, and image artifacts, which is hard because contrails evolve from thin bright lines when fresh into diffuse patches indistinguishable from natural cirrus within tens of minutes. Tracking must follow each contrail as it drifts with the wind, spreads laterally, fragments behind clouds, and changes brightness with sun angle. Attribution must then determine which aircraft produced each detected contrail. None of these tasks is independent of the others: a missed detection in frame t leaves the tracker with no boundary to follow, and that track is permanently broken. A tracking swap at a crossing point delivers a wrongly identified track to the attribution algorithm, which then assigns the wrong flight silently, with no indication that an error occurred. Detection errors corrupt tracking; tracking errors corrupt attribution.

Existing approaches treat these as a sequential pipeline (Chevallier et al., 2023; Sarna et al., 2025). A detector segments contrail pixels from visual appearance. A tracker links masks across frames by appearance similarity. An attribution algorithm then matches tracked shapes to advected flight paths. Each stage takes only its own inputs and passes results forward.

This one-way architecture has two structural problems. First, errors compound without recovery: a detection miss cannot be reversed by the tracker, and a tracking swap cannot be corrected by attribution. There are no feedback paths. Second, flight data enters the pipeline too late. Aircraft positions and wind fields are available from the start and could constrain the detector to search only where flights occurred, yet a sequential pipeline withholds this information until the final attribution stage, after detection and tracking are already done.

We address both problems by inverting the pipeline. Rather than detecting contrails and then attributing them, we start from flights and ask, for each one: does this flight's contrail appear in the image? Aircraft broadcast their positions via ADS-B (Automatic Dependent Surveillance-Broadcast). We combine these positions with wind and temperature fields from the ERA5 meteorological reanalysis (Hersbach et al., 2020) and run the CoCiP contrail lifecycle model (Schumann, 2012) to predict where each flight's contrail should appear in the camera image. These predictions become spatial prompt masks that tell a video segmentation model where to look. The model, SAM2 (Ravi et al., 2024) adapted for dense per-frame prompting, either confirms the contrail with a refined segmentation mask or rejects the prompt with low confidence. The shift is from open-world

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

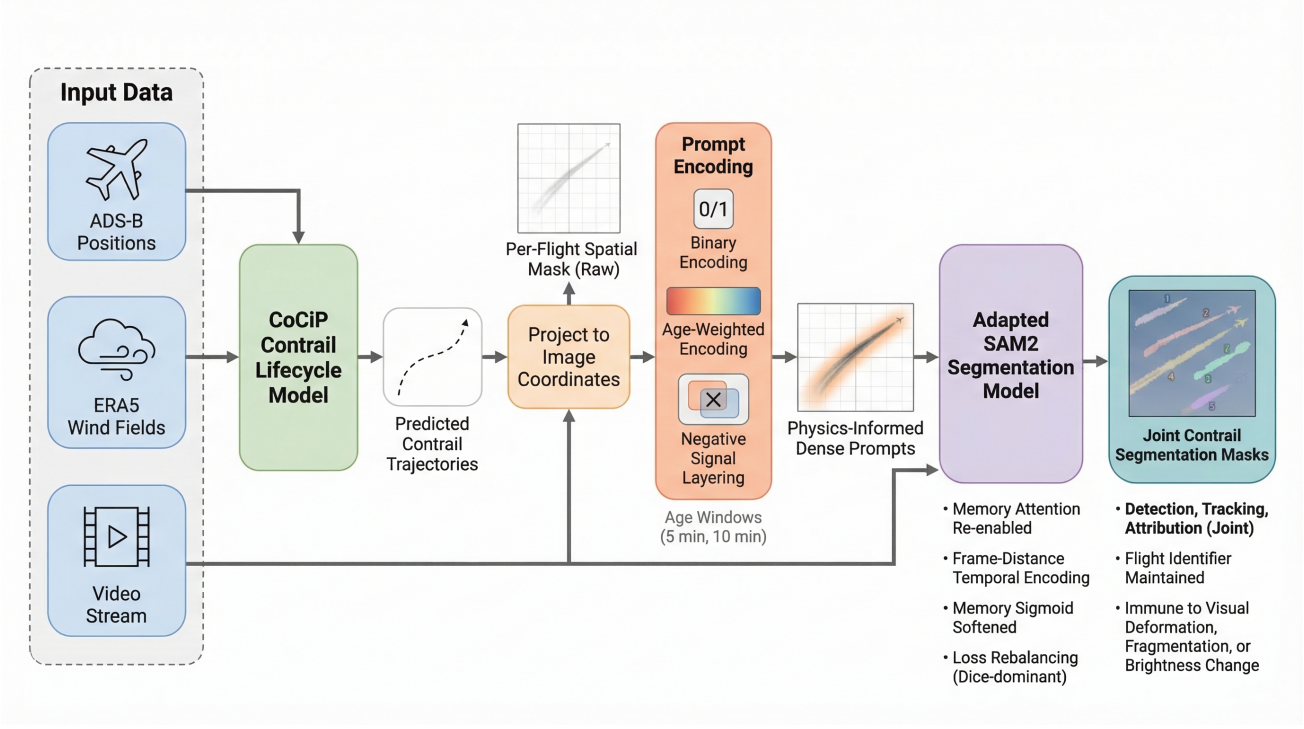

Figure 1. System overview. ADS-B positions and ERA5 wind fields feed the CoCiP contrail lifecycle model, whose predicted trajectories are projected into image coordinates to produce per-flight spatial masks. Three prompt encodings are evaluated: binary (1-minute window), and age-weighted with or without negative-signal layering (5- and 10-minute windows). An adapted Segment Anything Model 2 (SAM2) receives these dense per-frame prompts and outputs per-flight segmentation masks. The listed adaptations (memory attention on conditioning frames, frame-distance temporal encoding, a softened memory sigmoid, and dice-dominant loss rebalancing) are four of the five described in Sect. 4; the fifth, disabling the mask-decoder bypass, acts inside the SAM2 block and is not drawn separately.

detection (find all contrails anywhere in the scene) to constrained verification: physics specifies where to look, and the model decides what is actually there.

All three tasks follow directly. Detection is flight-aware: without a flight prompt, no output is generated for a location regardless of visual resemblance to a contrail. Tracking is implicit: each prompt carries a flight identifier across every frame, so no appearance-based association is needed; prompts follow atmospheric trajectories that remain geometrically consistent even when contrails deform, fragment, or cross. Attribution is by construction: each output mask corresponds to exactly one flight prompt.

A critical design choice underlies this inversion: we deliberately generate prompts for *all* flights, including those for which the meteorological model predicts no contrail. Rather than filtering flights on atmospheric pre-conditions, we let the visual model decide. The system must therefore learn both confirmation ("the contrail is here, refine its boundary") and rejection ("no contrail is visible despite the prompt, output nothing"), a capability we call *prompt rejection*. Prompt rejection is what makes

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

the system safe to deploy without a thermodynamic pre-filter: uncertainty in ERA5 humidity is absorbed by the visual model instead of propagating as silent false rejections. Beyond the pipeline structure, we investigate how much physical information each prompt should encode, ranging from flat binary presence masks to age-weighted encodings with suppression from competing flights. Sect. 3 details the encoding strategies; Fig. 1 shows the full pipeline. Our contributions are:

1. A framework that unifies contrail detection, tracking, and attribution into a single dense-prompting video segmentation task, with targeted modifications to SAM2's memory, temporal encoding, and loss for per-frame physics-derived prompts.
2. A systematic comparison of prompt design along two axes: how much information each prompt pixel carries, and how far back in the contrail's life the prompt reaches.
3. An evaluation protocol that goes beyond standard segmentation metrics to assess the full detection–tracking–attribution chain, covering pixel-level quality across IoU thresholds, temporal tracking continuity and fragmentation, prompt discrimination, flight-level attribution precision, and cross-contamination between overlapping flights.
4. A cluster-bootstrap uncertainty analysis that separates the effects we can defend from those we cannot. The attribution and segmentation-quality gains survive video-level resampling; the apparent detection-rate differences against the binary baseline do not.

2 Background and related work

This section covers contrail formation physics, surveys observational methods from satellite systems to ground cameras, and reviews related work on video object segmentation and contrail attribution.

2.1 Contrail formation and climate impact

When jet engines burn kerosene at cruise altitude, the hot exhaust mixes with cold ambient air, and the rapid cooling condenses water onto soot particles as ice crystals. The Schmidt–Appleman criterion (Schmidt, 1941; Appleman, 1953; Schumann, 1996) defines the thermodynamic conditions under which contrails form. Persistence depends on humidity: above ice saturation, crystals grow and the contrail can last for hours, spreading through wind shear into cirrus sheets; below it, the contrail sublimates within minutes.

The Contrail Cirrus Prediction model (CoCiP) (Schumann, 2012) simulates the full lifecycle of individual contrails, including ice crystal microphysics, radiative effects, and atmospheric dispersion. CoCiP divides each flight path into short segments and tracks each one forward in time: wind carries the segment's centreline, shear spreads it laterally, and microphysics governs ice crystal growth or sublimation. At any given moment, each segment has a known position, orientation, length, and width. We use this output to generate spatial prompts: the predicted contrail geometry is projected into camera image coordinates, producing a pixel-level map of where the contrail should appear. Sect. 3 details the projection and encoding.

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

2.2 Contrail observation

Contrails can be observed from satellites or ground cameras, each with distinct tradeoffs. Satellites cover wide areas and provide continuous monitoring (Mannstein et al., 1999), but young contrails are narrow and may fall below satellite pixel resolution. By the time contrails spread enough to be reliably detected from space, they have drifted far from the source aircraft, making attribution difficult because multiple flight paths intersect the contrail's current position. Machine learning has improved satellite detection (Ng et al., 2023), but attribution remains challenging at satellite resolution.

Ground cameras (Low et al., 2025; Van Huffel et al., 2025; Gourgue et al., 2025; Jarry et al., 2026) capture sky images at higher spatial resolution and regular intervals, showing contrails shortly after formation when they remain near the source flight and attribution is more straightforward. The tradeoff is limited geographic coverage, weather dependency, and a fixed field of view. We use ground camera data because it provides unambiguous attribution ground truth: the camera catches contrails within minutes of formation, while the aircraft is still in or near the field of view, so the association between contrail and flight is geometrically unambiguous. Annotated ground-camera collections have appeared recently on both sides of this tradeoff, from single-image segmentation corpora (Gourgue et al., 2025) to instance-level video sequences. The Ground Visible Camera Contrail Sequences (GVCCS) dataset (Jarry et al., 2026) provides the data for this work. It consists of sky image sequences captured at 30-second intervals from an all-sky camera at Brétigny-sur-Orge, France, beneath European routes. Each contrail instance is annotated with multi-polygon masks that follow the contrail's shape across frames, linked to a unique flight identifier obtained from ADS-B records. The dataset is partitioned into training and test splits at the sequence level. Our test set contains 24 videos (4,781 frames). CoCiP generates prompts for the 3,432 flight–video pairs that crossed the camera's field of view during these sequences. Under the 5-minute age-weighted window, 2,609 of them yield a non-empty in-frame prompt raster and are therefore presented to the model; 610 of those produced visible contrails that were annotated with ground truth masks, and the remaining 1,999 have no visible contrail and serve as negative examples that the model must learn to reject. Every annotated contrail in the test set received a prompt, so the prompting stage excludes no ground truth from the evaluation. Sect. 4.4 provides further details on the training protocol.

2.3 Segmentation and video object segmentation

Instance segmentation has progressed from encoder–decoder networks (Ronneberger et al., 2015) to region-proposal methods (He et al., 2017) and then to transformer architectures such as Mask2Former (Cheng et al., 2021b, a). Mask2Former has been applied directly to video contrail segmentation (Jarry et al., 2026), where it reaches strong per-frame accuracy but has no built-in attribution. In video object segmentation, memory-augmented models such as XMem (Cheng and Schwing, 2022), Cutie (Cheng et al., 2024), and DeAOT (Yang et al., 2022) propagate one or a few annotated frames through a video; none is designed to incorporate external domain knowledge that specifies object locations on every frame.

The Segment Anything Model (SAM) (Kirillov et al., 2023) introduced promptable segmentation, and SAM2 (Ravi et al., 2024) extends it to video with a temporal memory bank: a vision backbone (Hiera (Ryali et al., 2023)) extracts image features, a prompt encoder embeds the user's spatial input, and a mask decoder combines both with memory to produce segmentations.

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

In standard use, a human annotates one or two *conditioning frames* and SAM2 propagates through the rest. Recent adaptations cover camouflaged object segmentation (Zhou et al., 2025), semantic segmentation (Syed Ariff et al., 2025), and more (Zhang and Tang, 2025). Non-visual sensors have also been brought into the SAM2 family, but as additional *observation* streams to be segmented: SAM4D (Xu et al., 2025), for instance, jointly segments camera and LiDAR data by lifting both into a shared positional encoding. Our setting is the opposite. The non-visual input is never segmented and never observed at all; it is a forward prediction of where an object should be, supplied as a prompt on every frame. We are not aware of prior work that drives SAM2 this way. The reversal of the usual sparse-annotation ratio creates the problems that the modifications in Sect. 4 address.

2.4 Contrail attribution

Attribution algorithms match detected contrails to flights using geometric criteria (Chevallier et al., 2023; Riggi-Carrolo et al., 2023; Geraedts et al., 2024; Dalmau et al., 2025): trajectories are projected into image coordinates, advected by wind, compared with detected shapes, and matched by solving an assignment problem. This has been done on geostationary satellite data (Chevallier et al., 2023; Geraedts et al., 2024), with probabilistic treatments of uncertainty (Riggi-Carrolo et al., 2023), and from ground cameras in our previous framework (Dalmau et al., 2025). Such algorithms are now load-bearing: the airline avoidance trials cited in Sect. 1 measure their headline effect through automated satellite attribution (Sonabend-W et al., 2024; Sankar et al., 2026), and a recent benchmark quantifies how much the choice of attribution algorithm moves those numbers (Sarna et al., 2025). A common assumption across this literature, our own work included, is that detection and tracking are perfect, so the attribution algorithm receives clean masks and unbroken tracks. In practice they are not, and errors in either stage propagate silently. The present paper removes this assumption by evaluating the complete pipeline from raw images to flight-level attribution.

All post-hoc methods share the structural limitation described in the introduction: flight information enters too late to help with segmentation, and errors propagate unchanged through the pipeline. Our approach eliminates this by making attribution intrinsic to the segmentation process.

3 Flight-conditioned prompting

Every contrail traces back to an aircraft and a wind transport path. If we know where a flight occurred and how the wind carried the exhaust, we can predict where the resulting contrail should appear in the camera image. We exploit this to construct a spatial prompt for each flight on each frame: a pixel-level prediction of where that flight's contrail should lie, meaning the exhaust plume displaced by wind rather than the aircraft itself. How tight those prompts can be depends on how accurately we know the wind.

Prompt design is built up incrementally along two axes. The *age window* $\tau_{\max}$ sets how much of the contrail lifecycle to include: shorter windows give tighter localization, longer windows cover more of the visible track at lower spatial precision. The *encoding richness* controls what each prompt pixel carries. The simplest encoding is binary (presence only). Age-weighted

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

encoding adds a temporal gradient, so intensity encodes contrail freshness. On top of the age-weighted encoding, a negative signal from competing flights can be layered. These last two axes, negative signal (absent vs. present) and age window, are orthogonal by construction: the negative signal formula applies at any window width, and the window choice does not alter the signal representation. They form the 2×2 grid that Sect. 5 evaluates. The binary baseline sits outside that grid and differs from it on more than one factor at once, a limitation we return to when interpreting the comparison.

Fig. 2 illustrates the three encoding strategies on a scene with overlapping contrails. The window tradeoff is discussed under each encoding below; Sect. 5 evaluates windows of 5 and 10 minutes.

3.1 From flights to spatial prompts

Modern commercial aircraft broadcast their position continuously via ADS-B, transmitting GPS coordinates, barometric altitude, velocity vector, and a unique 24-bit ICAO identifier. Ground-based receivers collect these broadcasts into openly accessible databases such as the OpenSky Network (Schäfer et al., 2014), providing flight trajectory data for commercial air traffic.

We use `pycontrails` (Shapiro et al., 2024) to run CoCiP (Sect. 2.1) on each flight's ADS-B trajectory with ERA5 meteorological fields (Hersbach et al., 2020). The output is a time series of contrail segment states: centreline position, along-track length, cross-track width, and depth.

Meteorological inputs come from two ERA5 products (Hersbach et al., 2020). Wind, temperature, specific humidity, vertical velocity, and cloud ice water content are retrieved on model levels covering the cruise altitude range. Contrails spread sideways because wind speed changes with altitude (wind shear), so getting the vertical wind profile right matters more than horizontal detail. Model levels are spaced more finely in altitude at cruise height than the standard pressure-level product. Radiation variables (surface shortwave and longwave fluxes) are retrieved separately. Both products have hourly temporal resolution, linearly interpolated to each segment's time. Table 1 lists the CoCiP configuration parameters.

CoCiP transports each segment's centreline with the local wind velocity and models width growth through wake vortex dynamics in the first few minutes after formation, then diffusive spreading driven by atmospheric turbulence and wind shear. Aircraft-specific parameters (fuel flow, emissions index) are estimated by the Poll-Schumann performance model (Poll and Schumann, 2021). Integration uses a 30-second timestep to resolve the wake vortex phase, which shapes the contrail's initial geometry.

Two choices are critical for prompt generation and deserve particular attention. First, we deliberately disable both the Schmidt-Appleman formation filter and the initial-persistence filter. Standard CoCiP usage applies these filters to skip flights that are thermodynamically unlikely to form contrails, but ERA5 humidity at cruise altitude carries substantial uncertainty: the filters would reject some flights that do form visible contrails and pass some that do not. By running CoCiP on *all* flights and generating prompts regardless of the meteorological prediction, we let the visual model decide whether a contrail is actually present. This design choice is what makes prompt rejection possible (Sect. 5): rather than relying on a meteorological pre-filter that can fail silently, the model learns to output low confidence when visual evidence is absent, providing an explicit signal of uncertainty. Second, we correct ERA5 humidity bias using histogram matching (Shapiro et al., 2024) rather than a con-

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

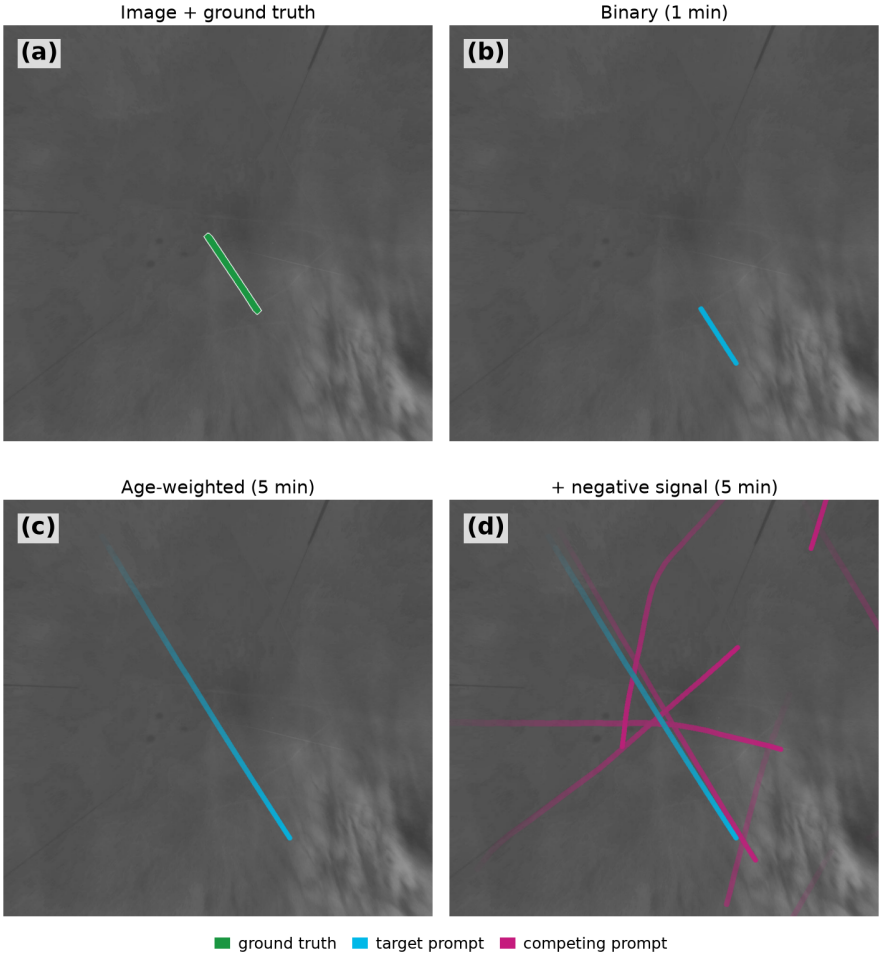

Figure 2. The three prompt encodings on a real test-set flight (video 5, frame 104). Sky images are shown desaturated throughout the paper so that the saturated overlay colours remain legible. (a) Camera image with the ground truth annotation (green, thickened for visibility). (b) Binary encoding: a flat presence mask restricted to the last minute of predicted geometry. (c) Age-weighted encoding over a 5-minute window, with intensity fading from bright (fresh, spatially reliable) to dim (old, potentially wind-drifted). The prompt extends well beyond the annotated contrail, which is expected: prompts are hints, not ground truth. (d) Adding the negative signal. Competing flights' predicted contrails (magenta) are layered as suppression, and where they cross the target's own prompt (cyan) the positive signal takes priority.

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

Table 1. CoCiP configuration used for prompt generation, implemented with `pycontrails` (Shapiro et al., 2024).

Parameter	Value	Description
Integration		
Integration timestep	30 s	Captures wake vortex descent in first minutes
Max contrail age	Video duration	Tracks contrails for the full observation period
SAC filter	Disabled	All flights prompted; model learns to reject non-contrailing flights
Persistence filter	Disabled	Same rationale; ERA5 humidity too uncertain to pre-filter
Meteorology & aircraft		
Humidity scaling	Histogram matching (Shapiro et al., 2024)	Distribution-corrected ERA5 humidity
Performance model	PSFlight (Poll and Schumann, 2021)	Fuel flow and emissions from flight parameters

stant scaling factor, which adjusts the full humidity distribution to match radiosonde observations and produces more realistic contrail width and persistence estimates.

We convert CoCiP segment descriptions into two-dimensional image coordinates using the camera projection model described in (Jarry et al., 2026), with known intrinsic and extrinsic parameters calibrated for the all-sky camera. Each segment's spatial extent, a polygon in three-dimensional world coordinates, is projected into image coordinates to produce a prompt mask. The mask for a given flight is the union of all projected polygons from that flight's active segments.

3.2 Prompt encoding strategies

We evaluate three prompt encodings of increasing richness: binary, age-weighted, and age-weighted with a negative signal. All three use the same projected contrail geometry; they differ only in what information each pixel carries.

3.2.1 Binary prompts

The simplest strategy is a binary mask: pixels inside predicted contrail polygons are set to 1, all others to 0. This matches SAM2's original mask prompt interface, where any nonzero region means "look here." Binary prompts tell the model where a contrail should be but carry no information about how old the contrail is or how confident the prediction is.

For each video frame at time t, we restrict prompts to contrail geometry younger than a maximum age $\tau_{\max}$. Three factors motivate a short window. First, wind errors accumulate over time: a prompt that is off by a few pixels at 1 minute may be off by tens of pixels at 10 minutes. Second, young contrails are thin, so prompt regions tightly match the real contrail width; older contrails have spread, producing wider prompts that give the model less spatial guidance. Third, fresh contrails are bright and high-contrast; older contrails fade, so prompting on young geometry aligns the spatial prior with the phase where visual evidence is strongest. A short window has drawbacks too. It may miss contrails that become visible only after a delay, for instance when clouds initially obscure formation and the contrail emerges minutes later, and it cannot capture the full curvature of a contrail that has spread along a long wind-sheared path. A longer window covers more of the contrail's visible extent and

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

catches late-appearing segments, at the cost of including older geometry that may have drifted from the actual contrail position. Sect. 5 evaluates 5- and 10-minute windows for the age-weighted encodings; the binary baseline uses a 1-minute window.

3.2.2 Age-weighted prompts

Binary prompts discard temporal information that could help the model. A contrail segment formed 30 seconds ago is visually bright and spatially well-localized; one formed 8 minutes ago is faint and may have drifted from the predicted position. We encode this by setting prompt pixel intensity proportional to contrail freshness:

$$p(x,y) = 1 - \frac{a(x,y)}{\tau_{\max}}, \tag{1}$$

where $a(x,y)$ is the age of the contrail segment covering pixel (x,y) in minutes. A segment that just formed has intensity 1.0; intensity decreases linearly with age, and segments at or beyond the age limit $\tau_{\max}$ contribute no prompt signal. This encoding gives the model a prior on where the contrail should be brightest and most reliably localized, effectively encoding temporal confidence into the spatial signal.

The age-weighted encoding interacts with the window tradeoff described above: with a wider window, the gradient from bright (young, reliable) to dim (old, uncertain) lets the model discount the spatial precision of older segments rather than treating all prompt pixels equally.

3.2.3 Prompts with a negative signal

When multiple flights produce nearby or overlapping contrails, the model may incorrectly segment one flight's contrail under another flight's prompt. Binary and age-weighted prompts encode only the target flight's predicted contrail location; they carry no information about other flights nearby. Adding a negative signal extends the age-weighted encoding by layering suppression from all other flights' prompts on top of the target flight's positive age-weighted prompt.

For each frame, we compute a union mask $u(x,y)$ that combines the age-weighted prompts of all flights in the scene. The prompt with negative signal for flight k is:

$$p_k(x,y) = \begin{cases} p_k^+(x,y) & \text{if } p_k^+(x,y) > 0 \\ -u(x,y) & \text{if } p_k^+(x,y) = 0 \text{ and } u(x,y) > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where p_k^+ is the age-weighted prompt for flight k (Eq. 1). Positive values indicate "your contrail should be here, with this confidence." Negative values indicate "another flight's contrail is here, not yours." Zero means background. The positive signal has priority: where the target flight's prompt overlaps with others, the target flight wins. The negative signal is also age-weighted, so the model learns that a negative region with a fresh nearby contrail deserves stronger suppression than one with only old, diffuse contrails. In principle, the negative signal could also be binary (present/absent) rather than age-weighted; we tested only the age-weighted variant, which preserves the temporal gradient on both sides. This means the negative signal axis

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

is not fully factorial: we did not train a binary-encoding + negative-signal variant, so we cannot isolate the negative signal's effect from the age-weighting with certainty.

The resulting prompt values lie in $[-1, 1]$ and are passed directly to SAM2's prompt encoder without rescaling. The prompt encoder's convolutional layers learn to map this range during fine-tuning: positive values activate the target flight's contrail region, negative values suppress competing flights, and zero represents uninformative background.

3.3 Edge cases

Several situations require care. When no contrail formed, either because conditions never satisfied the formation criterion or because it dissipated immediately, we still generate a prompt, since formation prediction from ERA5 humidity is uncertain; the model learns to reject such prompts when no visual evidence supports them. When turbulence fragments a contrail, the prompt shows the complete predicted trajectory while the model segments only the visible pieces; identity is preserved because the prompt carries the flight identifier regardless of fragmentation.

3.3.1 Prompts are hints, not hard constraints

The model does not segment exactly where the prompt says. Prompts tell the model approximately where to look, but the final segmentation boundary follows visual evidence in the image. In training, the ground truth mask is typically offset by several pixels from the prompt polygon, because wind errors, projection inaccuracies, and the contrail's actual shape all contribute, so the model learns that prompts and ground truth do not align perfectly. At inference, it can deviate from the prompt when the image shows the contrail is slightly elsewhere, rather than blindly tracing the prompt boundary. The age-weighted encoding reinforces this: bright (young) prompt regions are spatially reliable and the model trusts them closely, while dim (old) regions may be displaced, and the model relies more on what it sees. Without this tolerance for prompt–contrail misalignment, the system would fail whenever ERA5 wind estimates are off by even a few metres per second.

3.3.2 Unprompted contrails

Contrails that formed outside the camera's field of view and drifted into the observable region have no associated flight in our ADS-B data and are missed by design. This is not a practical limitation in the envisaged deployment. A network of cameras, each covering a different region of sky, would ensure that every contrail is young and promptable in at least one camera's field of view: the same contrail that appears as an old, unprompted feature in one camera was freshly formed and correctly attributed in another. Within each camera, our prompt-based model handles attribution for flights that cross the field of view, while a complementary open-world detection model (e.g., Mask2Former (Jarry et al., 2026)) could flag contrail-like features that lack a prompt as a consistency check.

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

4 Model architecture and training

We build on SAM2.1 (Ravi et al., 2024) with the Hiera Base+ backbone (Ryali et al., 2023), pretrained on SA-V. Standard SAM2 expects a human to annotate one or two frames, then propagates the segmentation through the video using a memory bank. Replacing those annotations with physics-derived mask prompts on every frame violates SAM2 defaults in three places, namely the training loop, the memory mechanism, and the loss function. We modify each below.

4.1 Dense prompt training loop

SAM2's training loop is designed for sparse human annotations: it samples one or two initial conditioning frames where the user provides a click or mask, then propagates to the remaining frames using predicted masks alone. Conditioning frames receive the prompt directly and skip the memory-based decoder; propagation frames rely entirely on temporal memory from previous predictions. This sparse-annotation assumption runs deep: it determines how the memory bank is populated, which frames query temporal context, and how the decoder is invoked. Replacing sparse human annotations with dense physics-derived prompts on every frame breaks it entirely. There are no propagation frames, every frame is a conditioning frame, and the training loop must be restructured accordingly.

We replace the training loop with sequential processing. At each frame t, the pipeline (1) encodes the image through Hiera B+ to extract multi-scale visual features at resolutions from $1/4$ to $1/32$, (2) encodes the flight's prompt mask into a dense feature embedding through the prompt encoder, (3) queries the memory bank via cross-attention to retrieve temporal context from previous frames, and (4) decodes a segmentation mask and object score through the mask decoder. The resulting features and predicted mask are stored in the memory bank before advancing to frame $t+1$. This means every subsequent frame can attend to all previous predictions, building up temporal context incrementally rather than propagating from a single reference.

We also disable a default where mask inputs bypass the decoder. In interactive use, the user's mask is already close to the ground truth and SAM2 copies it directly to the output, skipping the decoder for efficiency. Our prompt masks define where to look, not the exact object boundary. They are coarse physics predictions based on advection modelling, typically offset by a few pixels from the actual contrail edge and smoothed by the polygon rasterization process. Forcing every frame through the full mask decoder ensures that the final segmentation follows actual image evidence (the contrail's visual boundary) rather than the prompt geometry (the advection model's predicted polygon).

4.2 Memory adaptations

Three SAM2 defaults assume sparse interactive conditioning and fail under dense prompting.

4.2.1 Memory attention for conditioning frames

Standard SAM2 skips memory attention for conditioning frames, since the user already provided a correct mask and temporal context is unnecessary. When every frame is a conditioning frame, no frame queries memory and temporal propagation stops entirely. We re-enable memory attention for conditioning frames, limiting it to the 5 most recent to keep computation tractable

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

on long sequences. This allows each frame to benefit from context about what the model has seen and predicted in previous frames, even when a fresh prompt is available.

4.2.2 Temporal position encoding

Standard SAM2 assigns all conditioning frames temporal position zero, treating them as timeless reference anchors. With one or two reference frames among many propagation frames, this is fine. With dense conditioning, every frame gets position zero and the model cannot distinguish recent from distant context. We assign positions based on actual frame distance, so the model knows whether a memory entry is from one frame ago or twenty.

4.2.3 Memory sigmoid

SAM2 converts predicted mask logits to memory features through a sigmoid with scale 20 and bias -10. At scale 20, a logit of 0.3 maps to $\sigma(20\times0.3-10)=\sigma(-4)\approx0.02$, effectively zero. Memory is stored at 64×64 resolution ($1/16$ of the 1024×1024 input), so a contrail 2–4 pixels wide in the original image covers 0.1–0.25 memory pixels. Partial coverage produces logits near 0.1–0.3, which the hard sigmoid maps to near zero, erasing the contrail from temporal context before the next frame can query it. We use scale 1 and bias 0, which acts as a near-linear mapping near the origin, retaining partial values so the memory can encode "thin structure here at $\sim$20% confidence" rather than "nothing." Frames that cross-attend to this memory receive a valid, non-zero thin-structure signal from previous predictions.

The memory bank itself retains SAM2's default capacity (7 stored masks and 16 object pointers); only the cross-attention window over conditioning frames is restricted, to the 5 most recent. At the GVCCS temporal resolution of 30 seconds per frame, these 5 frames span 2.5 minutes of real time, enough to cover the period during which a contrail's appearance changes gradually and temporal context is most informative. Beyond this horizon, contrail morphology evolves substantially (spreading, fading, fragmenting), so older entries provide diminishing returns. The restriction also lowers GPU memory requirements, enabling training on longer sequences within hardware constraints. With more GPU memory, a wider attention window could help at higher frame rates, but 5 conditioning frames are sufficient at the GVCCS 30-second cadence.

4.3 Loss rebalancing for thin structures

The total loss per frame combines four terms:

$$\mathcal{L}=\lambda_{\text{dice}}\,\mathcal{L}_{\text{dice}}+\lambda_{\text{focal}}\,\mathcal{L}_{\text{focal}}+\mathcal{L}_{\text{IoU}}+\mathcal{L}_{\text{cls}} \tag{3}$$

where $\mathcal{L}_{\text{dice}}$ is the dice loss on the predicted mask, $\mathcal{L}_{\text{focal}}$ is the per-pixel focal loss (Lin et al., 2017), $\mathcal{L}_{\text{IoU}}$ is the L1 loss on predicted IoU, and $\mathcal{L}_{\text{cls}}$ is the focal loss on the object score.

Contrails are thin: a typical contrail covers less than 0.1% of the image pixels. SAM2's default loss weights focal at $\lambda_{\text{focal}}=20$ and dice at $\lambda_{\text{dice}}=1$. Focal loss sums over every pixel, so background pixels dominate the gradient. Dice loss normalizes by the mask area, giving thin objects and large objects equal weight. With the default weighting, the easiest way for the model to reduce total loss is to predict nothing: background focal loss drops to near zero, and the small dice penalty from missing a thin

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

contrail is negligible. We invert the weights to $\lambda_{\text{dice}} = 5$ and $\lambda_{\text{focal}} = 1$ (chosen by manual iteration on training convergence, not grid search), and set focal $\alpha = 0.75$ to up-weight the rare positive class. The area-normalized dice term now drives learning, with focal providing per-pixel boundary refinement rather than dominating the gradient.

Most flights crossing the camera do not produce visible contrails, so the model must learn to reject empty prompts. We supervise the object score head with focal loss ($\gamma = 2$, $\alpha = 0.75$), where $\alpha = 0.75$ up-weights the rare "object present" class. When the object score predicts absence, the corresponding mask loss is still computed but the target mask is empty, so the model learns both "nothing here" and "stop drawing pixels" from the same negative examples.

4.4 Dataset and training protocol

We use GVCCS (Jarry et al., 2026), the ground camera dataset described in Sect. 2.2, with instance-level contrail annotations linked to flight identifiers. The camera captures sky images at 1024×1024 resolution. Annotators labelled contrails with multi-polygon masks and unique identifiers persisting across frames. We partition at the sequence level to prevent information leakage.

We train without a held-out validation set, using all training sequences for learning. The dataset is small enough that holding out validation data would noticeably reduce the variety of contrail appearances and atmospheric conditions available for training. We instead select the number of training epochs per variant from training loss curves (Sect. 5.7). Early stopping on a validation metric would be more principled; we accept the less formal criterion in exchange for using all available training data.

4.4.1 Single-object lifecycle training

Rather than training on clips with multiple overlapping flights, we train on one flight at a time. Each training sample follows a single flight's contrail through its full lifecycle: appearance, persistence, and disappearance. This simplifies learning because the model refines one prompt into one mask per sample without managing multi-object interactions. At inference time, SAM2's multi-object tracking processes multiple flights in parallel, but training benefits from reduced complexity.

4.4.2 Object-centric sampling

We treat each (video, flight) pair as a separate dataset entry. Within each epoch, every prompted flight is visited exactly once, regardless of how many frames it spans or whether the flight has ground truth masks. Many prompted flights have no visible contrail and therefore no ground truth, and the model learns to reject these empty prompts as negative examples. A video with 15 prompted flights produces 15 training samples per epoch. This ensures short-lived contrails are sampled as frequently as long-lived ones, and negative examples are represented alongside positive ones.

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

4.4.3 Lifecycle sampler

For each flight, sampling starts at its first prompted frame and takes every second frame, so 13 model frames cover 26 real frames (13 minutes): GPU memory limits sequence length, and the 30-second GVCCS cadence is smooth enough that a 60-second effective interval still captures the gradual evolution of contrail appearance. Three additional frames past the last evidence of the contrail teach the model to output empty masks once visual evidence has vanished.

4.4.4 Augmentation

All spatial transforms (random flips, affine rotations of $\pm 25°$) are applied with identical parameters to every frame in a sample, and to the prompt and ground truth masks, to preserve temporal and physical consistency. Colour jitter is applied in two stages: a consistent pass across frames simulates different sky conditions, and a smaller per-frame pass mimics frame-to-frame illumination shifts from auto-exposure or passing clouds.

We deliberately omit ImageNet normalization. SAM2's pretrained checkpoint expects images normalized by ImageNet statistics, but ground camera sky images have very different statistics (dominated by blue and grey). During fine-tuning, we let the model adapt its internal normalization to the contrail domain. At inference, we override the default normalization with zero mean and unit standard deviation to match.

4.4.5 Optimization

We fine-tune from the pretrained SAM2.1 checkpoint using AdamW (Loshchilov and Hutter, 2019) with weight decay 0.1 (disabled for biases and LayerNorm parameters). The learning rate follows a 5% linear warmup from 5×10^{-7} to 5×10^{-6}, then cosine decay. The image encoder uses a lower peak rate of 3×10^{-6} with layer-wise decay of 0.9 to preserve low-level features learned on the SA-V pretraining data. Training uses mixed-precision bfloat16, gradient clipping at L2 norm 0.1, and gradient accumulation over 4 micro-batches (effective batch size 4) on two NVIDIA RTX 6000 Ada Generation GPUs (48 GB each). Total training time ranges from ∼39 hours (binary, 10 epochs) to ∼48 hours (20 epochs for the 10-min variants). Training duration is tuned per variant based on convergence analysis (Sect. 5.7).

5 Experimental results

Five prompt variants are compared on the GVCCS test set (24 videos, 4,781 frames, 2,609 prompted flight–video pairs with a non-empty prompt raster, of which 610 have ground truth annotations). A binary baseline uses flat 1-minute masks. The remaining four variants all use age-weighted encoding and are arranged in a 2×2 grid: negative signal (absent vs. present) × age window (5 vs. 10 min). Architecture and training procedure are identical across variants except for the number of training epochs (Sect. 5.7); prompt design is the primary variable. Fig. 3 shows a test-set frame in which the best variant receives prompts for 12 flights, simultaneously tracks eight of them, and rejects the remaining four; numbered, colour-coded predictions make attribution directly legible with no post-hoc matching.

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

5.1 Evaluation metrics

Standard segmentation benchmarks report pixel-level accuracy on individual frames, but our system must also track contrails across time and attribute them to specific flights. We therefore evaluate along three complementary axes, segmentation, tracking, and attribution, each with metrics chosen to capture a distinct aspect of performance.

5.1.1 Segmentation quality

We measure pixel-level accuracy using COCO-style mean average precision (mAP) (Lin et al., 2014), computed with `pycocotools`: predictions are ranked by confidence and precision is integrated over recall levels at each IoU threshold. We compute mAP at IoU thresholds from 25% to 75% in steps of 5%. The lower bound of 25% (rather than the COCO-standard 50%) reflects the geometry of contrails: these are thin structures where even a 1–2 pixel boundary error can halve the IoU, so a 50% threshold would penalize correct detections that are slightly misaligned. We report $\text{mAP}_{25:75}$ (averaged across all thresholds in this range), mAP_{25} (permissive, measuring coarse detection), and mAP_{75} (strict, measuring boundary precision). Note that this custom range differs from the COCO-standard $\text{mAP}_{50:95}$; the lower bounds reflect contrail geometry, and values are not comparable to standard COCO benchmarks.

5.1.2 Tracking

Four metrics assess temporal performance. *Detection rate* is the fraction of ground truth flights detected at least once (IoU $\geq$ 0.25 on any frame). It answers "did the system find this contrail at all?" *Track completeness* is the fraction of a flight's ground truth frames on which the model produces a detection, measuring how consistently the system follows the contrail across its lifetime. *Temporal IoU* is a flight's mean segmentation IoU across its ground truth frames. The reported aggregate averages the per-flight values over *all* ground truth flights, so flights that are never detected enter as zeros and the metric jointly reflects coverage and spatial accuracy. Because that conflation matters for interpretation, we also report temporal IoU restricted to flights detected at least once, which separates mask quality from coverage. *Fragmentation* counts distinct prediction identities covering each ground truth contrail: a value of 1.0 means the contrail was tracked under a single identity throughout, while higher values indicate identity breaks that would corrupt attribution. Fragmentation and completeness are likewise averaged over all ground truth flights, with undetected flights entering as zero, so both must be read alongside the detection rate rather than on their own (Sect. 5).

5.1.3 Attribution

We evaluate whether each detection is assigned to the correct flight. Predictions are greedily matched to ground truth annotations per frame by highest IoU (above 0.25). *Attribution precision* is the fraction of all matched predictions (IoU $\geq$ 0.25) that carry the correct flight identity, computed across all frames and all flights. Its denominator excludes flights the model never predicted on, so a variant can raise precision simply by predicting less; it must therefore be read against the detection rate and against recall, which is computed over all ground truth flights and cannot be inflated that way (Sect. 5). *Wrong attribu-*

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

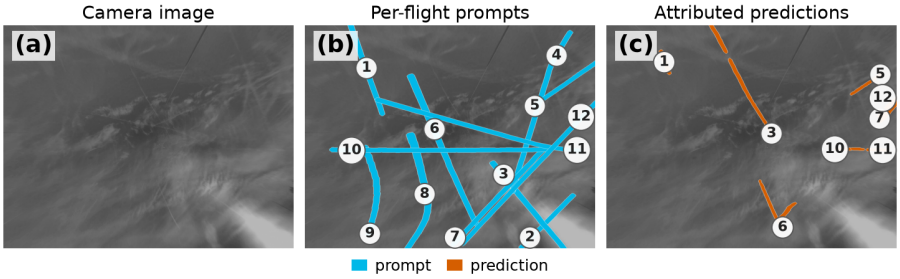

Figure 3. Attribution in action (video 2, frame 176, age-weighted with negative signal, 5-min variant). (a) Raw camera image with multiple simultaneous contrails. (b) Positive prompt footprints of the 12 prompted flights (cyan, thickened for visibility); numbered markers identify the flights (negative suppression signals are not shown for clarity). (c) Predictions (vermilion) carrying the matching flight numbers. Identity is established by construction: no post-hoc matching step is needed. Of the 12 prompted flights, 8 are confirmed with segmentation masks; flights 2, 4, 8, and 9 produce no prediction (prompt rejection), shown by the absence of their number in panel (c).

tion rate is the fraction matched to the wrong flight. Because no prediction in our test runs was matched to an unattributed (old, pre-existing) contrail annotation, precision and wrong-attribution rate sum to one. *Attribution recall* is evaluated at the first-detection event: for each ground truth flight, we ask whether the *first* frame on which the model produces a matched detection is attributed to the correct flight. This measures whether the system establishes the right identity from the outset, which matters because incorrect first-frame attribution would persist through the rest of the track. Attribution recall is the fraction of *all* ground truth flights whose first detection exists and is correctly attributed; it therefore equals the detection rate multiplied by the first-frame precision, and is bounded above by the detection rate. *Cross-contamination* measures how often a flight's predicted mask overlaps with another flight's ground truth, which can occur when nearby contrails bleed into each other's prompt regions. A contamination *event* is one predicted mask on one frame for one flight; high-contamination events are those where the overlap ratio exceeds 50%.

5.2 Prompt design comparison

Table 2 presents the results for all five variants.

5.2.1 The two grid axes are close to additive

One caveat applies to every comparison with binary: the binary baseline uses a 1-min window while the age-weighted variants use 5 or 10 min, so encoding richness and window width move together in that row and no experiment in this paper separates them. Within the 2×2 grid the two factors are properly crossed. The window axis is isolated by comparing 5-min against 10-min within each encoding block, and the encoding axis by comparing age-weighted against negative-signal within each window block. Age-weighted prompts raise $mAP_{25:75}$ from 0.303 (binary) to 0.369 at 5 min (+21.8%) and 0.348 at 10 min (+14.9%). Adding the negative signal pushes further, to 0.380 at 5 min (+3.0% over age-weighted) and 0.364 at 10 min (+4.6%), and the

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

Table 2. Prompt design comparison on the GVCCS test set (610 flights with ground truth, 1,999 negative-example flights). All variants use the same SAM2.1 Hiera B+ architecture. The four age-weighted variants form a 2×2 grid: negative signal (absent vs. present) × age window (5 vs. 10 min). Binary uses a flat 1-min mask without age weighting; its row therefore conflates encoding and window (Sect. 5). Training epochs are tuned per variant (Sect. 5.7). **Bold** marks the best value in each column; ties are bolded jointly. Metrics are defined in Sect. 5.1; note in particular that Precision is computed over matched predictions only, so it must be read against Detection and Recall. Video-level bootstrap 95% confidence intervals and paired best-minus-binary contrasts are in Fig. 5; the mAP columns have no interval (Sect. 7).

		Segmentation			Tracking				Attribution		
Encoding	**Window**	$mAP_{25:75}$	mAP_{25}	mAP_{75}	Detection	Completeness	Temp. IoU	Temp. IoU (detected)	Precision	Recall	Wrong
Binary (flat)	1 min	0.303	0.579	0.024	**93.4%**	0.724	0.500	0.535	89.0%	87.2%	11.0%
Age-weighted	5 min	0.369	0.666	**0.034**	91.3%	0.709	0.510	**0.559**	96.1%	88.5%	3.9%
	10 min	0.348	0.629	0.033	87.9%	0.642	0.489	0.556	**96.4%**	84.6%	**3.6%**
Age-wtd. + neg. signal	5 min	**0.380**	**0.693**	**0.034**	92.6%	**0.730**	**0.516**	0.557	96.2%	**90.3%**	3.8%
	10 min	0.364	0.657	0.033	90.0%	0.688	0.501	0.557	**96.4%**	87.2%	**3.6%**

best variant improves $mAP_{25:75}$ by 25.4% relative to binary. Shortening the window from 10 to 5 min helps at both encoding levels, by +6.0% for age-weighted and +4.4% with the negative signal. Within the grid these gains are close to additive, so the negative-signal and window factors interact only weakly.

Fig. 4 resolves the aggregate into its constituent IoU thresholds, which matters because $mAP_{25:75}$ averages over a range in which AP falls by more than an order of magnitude. Two things are visible that the single number hides. The absolute separation between the best variant and binary is widest at permissive thresholds, around 0.11–0.12 AP between 0.25 and 0.40, and shrinks steadily toward the strict end. The relative separation moves the other way, growing from 1.20× at IoU 0.25 to 1.42× at IoU 0.75. Both readings are correct, and they describe different things: most of the absolute gain comes from contrails that richer prompts delineate well enough to clear a loose threshold, while the growing ratio shows the improvement is not confined to the loose end. All five curves converge below AP 0.04 at IoU 0.75, so no variant delineates thin contrails to that tolerance with any regularity.

5.2.2 The 5-minute window resolves the detection–quality tradeoff

The 10-minute variants pay a detection-rate penalty: 87.9% (age-weighted) and 90.0% (with negative signal) against 93.4% for binary. We attribute this to prompt drift, since 10-minute-old geometry has accumulated enough ERA5 wind error that the prompt no longer sits on the contrail and the model treats the mismatch as evidence that no contrail is present. Figs. A1 and A2 show the mechanism on individual flights. The 5-minute variants recover most of this loss, reaching 91.3% (age-weighted) and 92.6% (with negative signal). Five minutes is short enough that ERA5 errors stay within the model's spatial tolerance, yet long enough for the age gradient and negative signal to improve boundary quality. Track completeness tells the same story: 0.730 with the negative signal at 5 min, against 0.688 at 10 min, 0.709 for age-weighted 5 min, and 0.724 for binary.

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

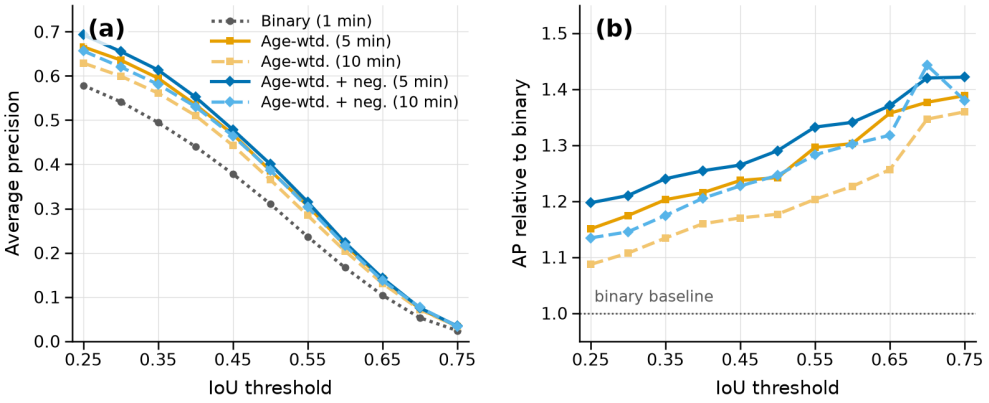

Figure 4. Average precision as a function of the IoU matching threshold, for all five variants. (a) Absolute AP. The ordering is stable across the whole range and every age-weighted variant sits above binary at every threshold. (b) AP divided by the binary baseline at the same threshold. The relative advantage of richer prompts grows monotonically with the threshold, from about 1.20× at IoU 0.25 to about 1.42× at IoU 0.75, even as the absolute gap narrows. Line style and marker encode the variant redundantly with colour so the figure survives greyscale reproduction. The $\text{mAP}_{25:75}$ column of Table 2 is the mean of each curve in panel (a).

5.2.3 Coverage and mask quality move separately

Temporal IoU aggregated over all ground truth flights (0.500 for binary, 0.516 for the best variant) mixes two different things, because undetected flights enter the average as zeros. Restricting the same statistic to flights the model actually detects separates them, and the separation is stark. Among detected flights, all four age-weighted variants are indistinguishable from one another (0.559, 0.556, 0.557, and 0.557 for age-weighted 5 and 10 min and negative-signal 5 and 10 min respectively), while binary sits below them at 0.535. Once a contrail is found, in other words, the window width and the negative signal make essentially no difference to how well its boundary is drawn; what they change is how often it is found and for how many frames it is held. The one encoding change that does improve per-frame mask quality is the age gradient itself, worth roughly 0.02 IoU over binary. This decomposition also explains why the aggregate temporal-IoU spread across the grid (0.489 to 0.516) is wider than the detected-only spread (0.556 to 0.559): almost all of it is coverage.

5.2.4 Attribution

All non-binary variants reach attribution precision above 96%, a substantial jump from 89.0% for binary. Without the age gradient the model confuses overlapping flights far more often, at 11.0% wrong attributions for binary against 3.6–3.9% for every other variant. The gain is flat across non-binary variants (96.1–96.4%) whether or not the negative signal is present and at either window, which suggests the age gradient alone accounts for most of it.

Precision on its own would be a weak claim here, because its denominator counts only predictions that matched a ground truth mask, so a variant can raise it by predicting less. The two axes have to be read together. Age-weighted 10-min posts the

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

Table 3. Identity quality across all five variants. Fragmentation is the mean number of prediction identities covering each ground truth contrail (1.0 = one identity per contrail). Flights that are never detected contribute 0, so values below 1.0 reflect missed detections rather than cleaner tracking and must be read against the detection rates in Table 2. Contamination events count (prediction, frame) pairs whose mask overlaps another flight's ground truth; high events are those with overlap ratio >50%. Raw event counts scale with how often a variant predicts at all, so the percentage in the last column, not the count, is the comparable quantity.

Encoding	Window	Fragm.	Contam. events	High (>50%)
Binary (flat)	1 min	1.19	13,614	2,402 (17.6%)
Age-weighted	5 min	1.00	8,560	687 (8.0%)
	10 min	0.96	8,227	613 (7.5%)
Age-wtd. + neg. signal	5 min	1.01	8,672	641 (7.4%)
	10 min	0.98	8,726	566 (6.5%)

joint-best precision (96.4%) at the worst detection rate in the study (87.9%), which is consistent with exactly that selection effect. The negative-signal 5-min variant posts statistically the same precision (96.2%) at a 92.6% detection rate, 4.7 points higher, and the highest attribution recall of any variant: 90.3%, against 84.6% for age-weighted 10-min and 87.2% for binary. Since recall is computed over *all* ground truth flights rather than over matched predictions, it cannot be inflated by predicting less, and it is the column that identifies the best variant. Sect. 6 explains the mechanisms behind both effects.

5.2.5 Identity quality

Table 3 extends two identity-centric metrics, fragmentation and cross-contamination, to all five variants. Binary tracking is worse on both. Its mean fragmentation of 1.19 indicates that roughly one in five contrails is split across multiple prediction identities, whereas every age-weighted variant stays at or near 1.0, and 17.6% of binary's overlap events are high-contamination (mask overlap ratio above 50%) against 6.5–8.0% for the richer encodings. Identity splits and pixel leakage across flights are precisely the errors that corrupt attribution in sequential pipelines, and the age gradient largely removes both, with the negative signal reducing high-contamination events further at each window size.

Two features of the table need care. The fragmentation values below 1.0 for the 10-minute variants (0.96 and 0.98) are not better tracking; they arise because flights that are never detected contribute zero, so the metric partly rewards missing contrails. Read against the detection rates in Table 2, the honest statement is that binary fragments tracks and no age-weighted variant does. Similarly, the raw contamination event counts are not comparable across variants, because a variant that predicts on more frames has more opportunities to contaminate; binary's 13,614 events partly reflect its higher detection rate. The high-contamination percentage in the last column normalizes for this and is the figure to compare.

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

5.3 How much does the visual model add? A prompt-as-prediction baseline

Because prompts already encode where physics expects each contrail, one may ask how much of the reported accuracy is contributed by SAM2 at all. We therefore evaluate a training-free baseline that treats the rasterized CoCiP prompt itself (binarized, 5-minute window) as the predicted mask on every ground truth frame. The physics prior alone is far from a usable segmentation. Across the 14,262 ground truth frames of the 610 annotated flights, the raw prompt achieves a mean frame IoU of 0.042 against the annotation (median 0.0) and reaches IoU ≥ 0.25 on only 4.6% of frames. Aggregated the same way as the model metrics, that is a mean temporal IoU of 0.047 per flight, with only 0.8% of flights attaining a mean IoU above 0.25. Three factors drive the gap: wind-advection error displaces the prompt from the visible contrail, the projected polygon is systematically wider or narrower than the observed plume, and prompts persist on frames where the contrail is not, or not yet, visible. On the matched per-flight statistic, the trained model lifts mean temporal IoU from 0.047 to 0.516, roughly an eleven-fold increase. The visual refinement therefore contributes essentially all of the segmentation accuracy; what the physics prior contributes is the localization and the identity, neither of which the segmentation metrics reward directly.

This baseline bounds the contribution of the prior from below. It says nothing about the upper end, that is, how a competing detect-then-attribute architecture would score under the same protocol. For a rough external anchor, the Mask2Former panoptic baseline published with GVCCS (Jarry et al., 2026) operates on the same sequences but is neither flight-conditioned nor evaluated on attribution, so its segmentation numbers are not commensurable with ours and we do not tabulate them. Sect. 7 discusses why a fair comparison requires a shared protocol that we have not yet built.

5.4 Uncertainty quantification

All metrics in Table 2 are point estimates on a single test set; to assess their stability we compute cluster bootstrap confidence intervals that respect the correlation structure of the data: entire videos (the 18 test sequences containing annotated flights) are resampled with replacement 10,000 times, and every per-flight or per-prediction record from a sampled video enters the pooled statistic. Fig. 5 shows the resulting 95% intervals for the four headline metrics.

The per-flight distributions explain why the aggregate completeness difference is small: most flights are tracked equally well by both variants (330 of 610 within ± 0.02), while 133 improve and 147 regress under the best variant, with large movements in both directions. That is offsetting redistribution, not the absence of an effect.

Three conclusions follow. First, the attribution-precision gap between binary and every age-weighted variant is decisively significant: the paired video-level bootstrap for the best variant minus binary gives $+7.1$ points with 95% CI $[+5.4, +8.9]$ and $p(\Delta \leq 0) < 10^{-4}$, and the same holds for temporal IoU ($+0.016$, CI $[+0.008, +0.024]$). Second, the within-grid effects are also robust. At the 5-minute window, adding the negative signal improves completeness ($+0.021$, CI $[+0.011, +0.030]$) and detection rate ($+0.013$, CI $[+0.005, +0.020]$); at fixed age-weighted encoding, shortening the window from 10 to 5 minutes improves completeness ($+0.067$), detection rate ($+0.035$), and temporal IoU ($+0.021$), all with $p(\Delta \leq 0) < 10^{-2}$. Third, the differences in detection rate (-0.8 points, CI $[-2.2, +0.8]$) and completeness ($+0.6$ points, CI $[-1.2, +2.5]$) between the binary baseline and the best variant are *not* statistically distinguishable. The binary baseline finds contrails about as often as the best

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

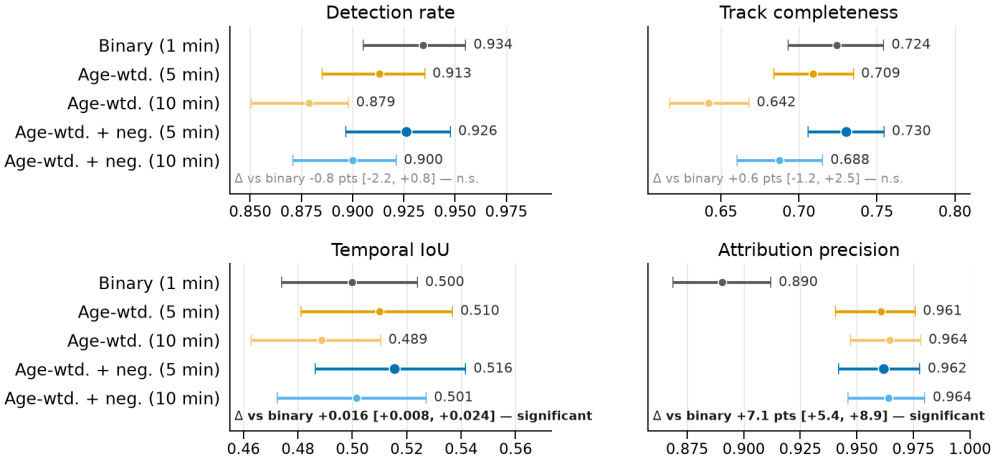

Figure 5. Headline metrics with 95% video-level cluster bootstrap confidence intervals (10,000 resamples of the 18 annotated test videos). Colour encodes the prompt family (grey: binary; orange: age-weighted; blue: age-weighted with negative signal); saturated marks denote 5-minute windows, and the enlarged marker is the best variant. The footer of each panel gives the *paired* best-minus-binary contrast, which is the quantity the significance claims rest on: the marginal intervals drawn in the panels are wide and overlapping because they carry between-video variance that cancels in the paired comparison. Attribution precision separates decisively, temporal IoU separates, and detection rate and track completeness do not.

variant; what it cannot do is delineate them as accurately or attribute them as reliably. We therefore claim the segmentation-quality and attribution gains as the robust contributions of richer prompts, and explicitly do not claim a gain in raw detection coverage.

The paired and marginal views differ enough to be worth stating plainly, because Fig. 5 shows both. The marginal intervals for detection rate and completeness overlap heavily across all five variants, yet several within-grid contrasts are significant under pairing. Nothing is inconsistent here: the marginal intervals carry between-video variance, and some test videos are simply harder than others for every variant, so that component cancels when the same resampled videos are scored under two variants at once. The paired contrast is the appropriate test for "does this design change help", and the marginal interval is the appropriate description of "how well would this variant do on a new set of videos from this site". The second is much wider than the first, and 18 videos is not many.

5.5 Visual analysis

5.5.1 Tracking sequence

Fig. 6 shows three frames sampled from a 25-frame tracking window (frames 205–229) in which the best variant (age-weighted with negative signal, 5 min) follows flight 34ba902a (video 14) through a scene dense with competing traffic. Under the binary baseline this flight reaches 40.9% completeness; the age-weighted with negative signal variant raises it to 56.8%. The prompt

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

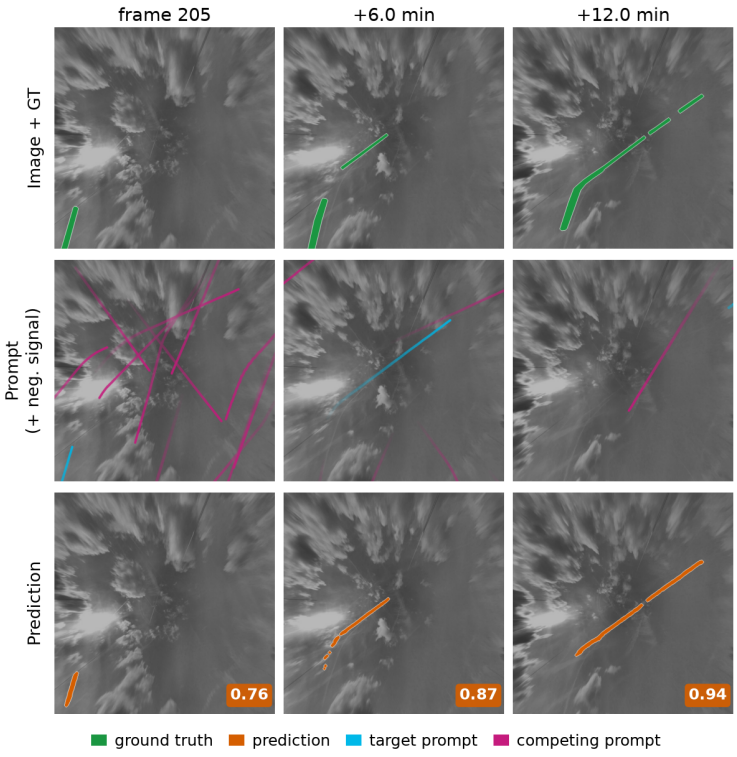

Figure 6. Tracking sequence for flight 34ba902a (video 14, age-weighted with negative signal, 5-min variant). The three columns are frames 205, 217 and 229, labelled by elapsed time at the 30-second GVCCS cadence. Row 1: image with ground truth overlay (green). Row 2: age-weighted prompt with negative signal overlay, where *cyan* marks the target flight's age-weighted footprint and *magenta* marks regions attributed to competing flights (the negative suppression signal). Row 3: prediction overlay (vermilion) with object score badge. On this flight the variant reaches 56.8% completeness against 40.9% for the binary baseline, with scores rising from 0.76 to 0.94 across the shown frames. In the third column the target's own prompt has almost left the frame while the prediction remains locked on the contrail, which is temporal memory rather than the prompt doing the work.

row makes the difficulty of the scene immediately visible: magenta regions indicate where competing contrails from other flights claim the image, while a narrow cyan band marks the target contrail's predicted position. The negative signal suppresses attention to competing structures and allows SAM2 to lock onto the correct trajectory, with prediction scores rising from 0.76 to 0.94 across the shown frames as the visual evidence accumulates in memory.

5.5.2 Both design axes on a single flight

The same contrast appears on individual flights across the whole 2×2 grid. Fig. A1 tracks one flight under all five variants: every variant fires confidently on most frames, but only the 5-minute age-weighted prompts keep the mask on the annotated

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

contrail, while the binary and 10-minute predictions sit displaced from it and fail the IoU ≥ 0.25 match. The difference is delineation, not discovery, which is the aggregate finding of Sect. 5.4 reproduced on a single case.

5.5.3 Age window trade-off

The optimal window depends on per-flight geometry and surrounding traffic density in ways the headline metrics do not capture, and the two failure directions are opposite. A contrail that becomes visible only several minutes after formation is served poorly by the 5-minute window, which covers just the recent tip, and well by the 10-minute window (Fig. A2). A crowded scene is the mirror case: the 10-minute window sweeps older crossing contrails into the negative signal, and the resulting suppression is aggressive enough to eliminate the detection entirely, where the 5-minute window avoids the contamination (Fig. A3). An adaptive window, calibrated to ERA5 wind uncertainty and local scene density, is the natural extension.

5.6 Prompt discrimination

Most prompted flights do not produce visible contrails in the GVCCS dataset, so the model must learn to say "no." We evaluate this through the object score, a scalar output indicating confidence that a contrail is present. For each prompted flight we take the maximum score over its video; a flight whose prompts never trigger any prediction receives a score of zero. Under the 5-minute window, 2,609 of the 3,432 prompted flight–video pairs produce at least one non-empty prompt raster; 610 of these have ground truth annotations (visible contrails) and 1,999 do not.

Fig. 7 shows the resulting score distributions and the receiver operating characteristic (ROC) of prompt rejection for the best variant. The inference pipeline records predictions only above its 0.5 decision threshold, so flights that never cross it appear at score zero. Rates at the 0.5 operating point are therefore exact, while the AUC treats sub-threshold flights as ties. The two populations are well separated: flights with a visible contrail reach a mean maximum score of 0.79 against 0.04 for empty prompts, and the area under the ROC curve is 0.969. At the operational threshold of 0.5, the model confirms 94.3% of flights with visible contrails (5.7% false-negative rate) while rejecting 93.5% of empty prompts (6.5% false-positive rate). Because 77% of the 2,609 prompts with a non-empty raster are negatives, this operating point matters. Accepting every prompt would report 2,609 contrails where the model reports about 705, so the rejection stage removes roughly three-quarters of the candidate set. The residual errors are faint contrails the model misses and coincidental contrail-like features under empty prompts, and the failure-mode analysis below quantifies both.

The video-level bootstrap interval for the AUC is $[0.943, 0.971]$, which brackets the point estimate very asymmetrically: the point estimate of 0.969 sits close to the upper limit. This is a property of the statistic rather than a computational error. AUC is bounded above by 1 and the sampling distribution is strongly left-skewed at values this high, so a percentile interval piles up near the ceiling. We report the interval as computed and caution against reading its width symmetrically; a bias-corrected and accelerated interval would be a better choice for this particular statistic, and we did not compute one.

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

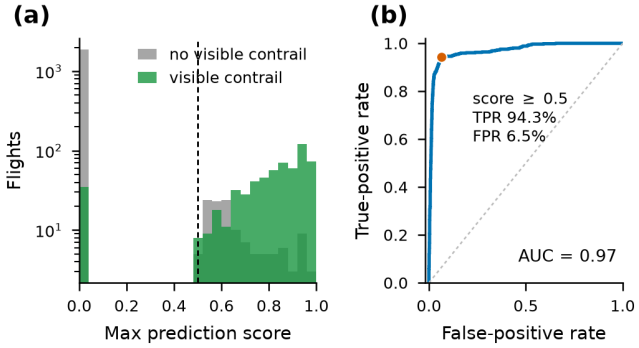

Figure 7. Prompt rejection by the best variant (age-weighted with negative signal, 5 min). (a) Distribution of per-flight maximum prediction scores (log-scale counts): flights with a visible (annotated) contrail in green, prompted flights without a visible contrail in grey; the dashed line marks the operational 0.5 threshold. (b) ROC curve of accepting a flight based on its maximum score (AUC 0.97); the marker shows the 0.5 operating point (true-positive rate 94.3%, false-positive rate 6.5%).

5.7 Training convergence

Age-weighted variants reach lower total loss (∼4.0) than binary (∼5.3) at comparable epochs. The richer continuous signal gives the model more to work with per frame, and the loss curve flattens accordingly. Binary prompts converge more slowly because the model must recover temporal context from visual features alone.

We therefore train each variant for a different number of epochs: 10 for binary, 15 for the 5-minute variants, 20 for the 10-minute variants. This per-variant tuning is the most serious confound in the study, and we want to state its extent precisely rather than argue it away. The richer variants receive 50 to 100% more training iterations than binary, so an unknown part of their advantage may come from longer training rather than from prompt design. Epochs were selected from training loss curves rather than test-set metrics, but the judgement that binary was overfitting past 10 epochs was confirmed on test videos, and that is a form of test-set selection. Sect. 4.4 explains why no validation split was held out; the cost of that decision is visible here.

We can bound the confound only by argument, not by experiment. Binary reaches its best performance at 10 epochs and degrades afterwards, while the age-weighted variants keep improving through 15 to 20 epochs, so equalizing the budget at 10 epochs would most likely widen the gap rather than close it. That reasoning is plausible and it is also unverified, because we did not evaluate the age-weighted variants at 10 epochs. The decisive control is to score every variant at a common epoch count; we leave it to follow-up work, and readers should treat the binary-to-best comparison as confounded until it is done. The within-grid comparisons are unaffected wherever the epoch count matches, which covers the negative-signal contrast at each window; only the window contrast (15 against 20 epochs) and the binary contrast carry the confound.

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

5.8 Analysis of the best variant

The previous subsections compared variants against each other. Here we examine the best-performing configuration (age-weighted with negative signal, 5 min) in more detail, focusing on what it does well and where it falls short.

5.8.1 Thin structures are hard

The gap between mAP_{25} (0.693) and mAP_{75} (0.034) needs context. mAP_{75} aggregates precision and recall over confidence thresholds at a strict IoU criterion, and it collapses whenever high-confidence predictions cluster below the 0.75 boundary. The low value does not imply that nearly all contrails are missed. It reflects that boundary delineation to better than IoU 0.75 is rare even when the contrail is detected and correctly located, which is what one expects for objects two to four pixels wide: a one-pixel boundary error on a two-pixel-wide object costs roughly a third of the IoU outright.

Performance on medium-sized objects ($mAP_{25:75} = 0.521$) is much higher than on small objects (0.371), which is consistent with a spatial-resolution bottleneck but does not by itself identify where in the network that bottleneck sits. Size-dependent AP of this kind appears in essentially every COCO-style evaluation and would be present in a single-frame model with no memory at all. We suspect the memory pathway contributes, since memory is stored at 64×64 (a sixteenth of the 1024×1024 input), so a two-pixel contrail occupies well under one memory cell and temporal context is correspondingly coarse. That remains a hypothesis: isolating it would require training a memory-free variant and comparing mAP_{75}, which we have not done.

5.8.2 Temporal consistency

Each ground truth contrail is covered by essentially a single prediction identity, the small excess above 1.0 reflecting rare frame-boundary effects where the last few frames of a fading contrail are picked up under a second short-lived prediction. This follows directly from the prompt-based architecture, in which each flight keeps its identity across all frames without appearance-based re-identification. A conventional VOS pipeline would fragment more, because identity has to be re-established at each frame through appearance matching, and appearance matching fails exactly when contrails spread, fade, or change brightness with sun angle. Prompt-based identity is indifferent to all three.

Track completeness of 0.730 means the model holds detection across roughly three-quarters of a contrail's visible lifetime on average. Detection also happens quickly. Among flights the best variant ever attributes correctly, the median delay between a contrail's first annotated frame and its first correct detection is zero frames, the upper quartile is one frame (30 s), and 96.6% are confirmed within two minutes of first visibility. That latency matters for the envisaged monitoring application, where feedback about an avoidance manoeuvre is most useful while the aircraft is still in the sector. The remaining quarter of frames are those where confidence drops below threshold, which happens in three situations: the contrail becomes very faint and the visual evidence weakens; wind advection error displaces the prompt far enough that the model cannot find the contrail in the prompted region; or appearance changes abruptly, for instance with sun angle, and the model briefly loses confidence before recovering from temporal memory.

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

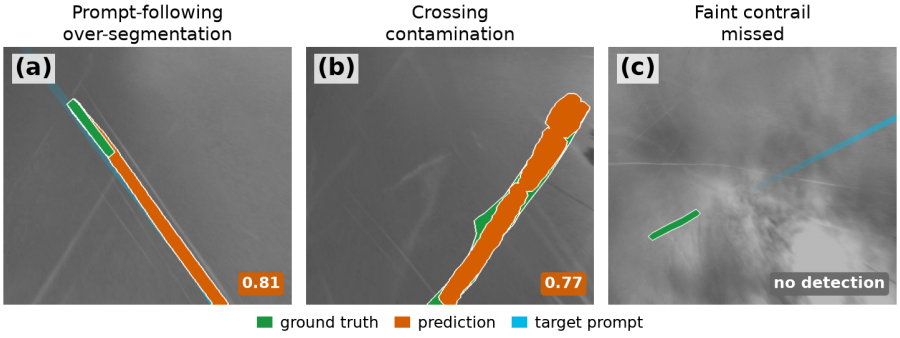

Figure 8. Three characteristic failure modes of the best variant (ground truth in green, target prompt in cyan, prediction in vermilion; masks thickened for visibility). (a) Prompt-following over-segmentation (flight 324b62ce, video 12): the wind-drifted prompt no longer matches the short visible contrail and the model traces the prompt instead, producing a confident mask (score 0.81) that covers the annotation but is far larger, so it fails the IoU match. (b) Crossing contamination (video 12, frame 97): the prediction shown belongs to flight 324b5124, while the green annotation it covers belongs to the neighbouring flight 324b65d8, 86% of which lies under the wrong flight's mask. Only the neighbour's annotation is drawn, so the panel shows a prediction that looks correct and is attributed to the wrong aircraft, which is precisely why this failure is hard to catch without per-flight ground truth. (c) Faint contrail missed (flight 33638739, video 5): the contrail is annotated (green) but the model, prompted along the drifted cyan band, never produces a detection.

5.8.3 Cross-contamination

When multiple flights have nearby prompts, the model may segment one flight's contrail under another flight's prompt (Table 3). The residual contamination is partly a data artifact rather than a model error: in dense airspace, ground truth masks from different flights naturally overlap because contrails spread and merge over time.

5.9 Failure modes

We manually categorized the 50 lowest-performing flights (by temporal IoU) from the best variant to understand where the system fails. Fig. 8 illustrates the three dominant modes on real test frames.

Wind estimation errors cause roughly one-third of failures. ERA5 wind fields miss mesoscale variability, and errors exceeding $\sim$5 m/s displace prompts by tens of pixels from the actual contrail. When displacement exceeds the model's spatial tolerance, the prompt provides misleading guidance and the model either rejects a real contrail, treating the mismatch as evidence of absence, or traces the drifted prompt itself and produces a confident mask far larger than the visible contrail, as in Fig. 8A.

Overlapping prompts cause roughly one-quarter of errors, concentrated in busy airspace where many flights produce nearby contrails. When prompt masks overlap substantially, even the negative signal cannot fully disambiguate boundary pixels. It tells the model that another flight exists, but when two contrails are separated by only a few pixels, correct boundary assignment is beyond the model's spatial resolution.

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

Unprompted contrails account for about one-fifth of missed detections. These are real contrails that drifted into the field of view from regions without ADS-B coverage, or from flights too far away for reliable advection modelling. These are missed by design: the system only examines prompted locations, and cannot attribute what it cannot identify.

The remaining errors are discrimination failures: accepting empty prompts (false positives from coincidental visual features) or rejecting real contrails (false negatives on faint or partially occluded contrails where the visual evidence does not meet the model's confidence threshold).

6 Discussion

Richer prompts and shorter age windows both help, and they do so for different reasons. This section traces the interaction between the two design axes and explains the attribution improvement.

6.1 Window and negative signal interact weakly

The negative signal gains most at 5 minutes because prompt accuracy and competing-flight suppression reinforce each other. At 5 min, ERA5 wind errors have had less time to accumulate, so both the positive and negative signals are spatially reliable. At 10 min, older segments have drifted further from the actual contrail, partially undermining both signals. That age-weighted prompts at 5 min score closer to negative-signal prompts at the same window than to age-weighted prompts at 10 min (Table 2) indicates that prompt accuracy matters more than the negative signal alone.

The detection-rate pattern carries the most practical weight. Both 10-minute variants fall below binary and both 5-minute variants close most of that gap, to the point where the remaining deficit is within resampling noise (Sect. 5.4). Five minutes is a sweet spot: wind errors remain within the model's spatial tolerance, and past 5 minutes accumulated drift starts causing false rejections. Accurate prompts do more than sharpen boundaries. They restore the detection reliability that wider windows give away.

6.2 Why richer prompts improve attribution

The jump in attribution precision is the single largest improvement from adding temporal structure to prompts, and it comes almost entirely from the age gradient. The binary model misattributes predictions mostly at flight crossings, where flat masks provide no spatial cue for disambiguation. A bright prompt region (young, well-localized) is a stronger spatial cue than a dim one (old, drifted), even when two flights overlap in space, and layering the negative signal on top adds almost nothing further to precision.

The negative signal matters for recall and contamination instead. Marking pixels as belonging to another flight suppresses missed detections at crossings, which is what lifts attribution recall, and high-contamination events drop accordingly because the signal limits pixel leakage across flight boundaries.

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

6.3 Computational cost

The system runs in three stages at inference. Prompt generation (flight advection and camera projection) runs on CPU and scales linearly with the number of active flights, taking approximately 0.3 s per frame for a typical scene of 10–20 active flights. SAM2 image encoding runs once per frame on GPU regardless of flight count (∼0.15 s on an RTX 6000 Ada). Memory attention and mask decoding run once per prompted flight per frame (∼5 ms per flight for decoding, plus per-object memory-bank updates). For a scene with 20 active flights, these components total well under one second per frame, more than an order of magnitude inside the 30-second inter-frame interval of the GVCCS camera. Because only the per-flight decoding scales with traffic, denser airspace (50+ flights) adds a fraction of a second and remains far within the inter-frame budget on current hardware.

6.4 Future directions

Multi-channel prompts could carry the meteorological context CoCiP already computes (the Schmidt–Appleman criterion, persistence probability, relative humidity over ice) at no labeling cost and with no architectural change. Adaptive age windows could tighten in calm conditions and widen in windy ones, automating the detection–quality tradeoff. Finally, combining prompt-based attribution with a conventional open-world detector would close the coverage gap for contrails from unknown or distant flights.

7 Limitations

The main limitation is wind accuracy. ERA5 reanalysis fields have limited horizontal resolution and hourly temporal resolution, missing mesoscale variability, jet stream gradients, and turbulence. When wind errors exceed roughly 5 m/s, prompts can be displaced by tens of pixels from the actual contrail. Higher-resolution regional models or data assimilation that refines wind estimates from observed contrail motion could help.

Dense traffic is another limitation. When many flights pass through similar airspace, prompts overlap and the model struggles with boundary assignment even with the negative signal. This limitation is geometric: at some density of overlapping contrails, no prompt encoding can disambiguate which pixels belong to which flight without sub-pixel spatial precision.

Statistical uncertainty is quantified for the tracking and attribution metrics through the video-level bootstrap of Sect. 5.4, but not for mAP, because resampling mAP means re-running the COCO evaluation inside every replicate. This is a real gap rather than a formality, since $\text{mAP}_{25:75}$ is the metric the abstract leads with and the only headline number with no interval attached. The within-grid mAP differences, 0.369 against 0.380 for instance, should therefore be read with caution; the ordering in Fig. 4 is consistent at every threshold, which is reassuring but is not a significance test. A full 10,000-replicate bootstrap is expensive; a delete-one-video jackknife over the annotated videos would give a usable standard error, and we leave it, together with the common-epoch control of Sect. 5.7, to follow-up work.

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

The test set, while densely annotated, comes from 24 sequences at a single site. Effective sample sizes for video-level resampling are modest, and the marginal intervals in Fig. 5 are correspondingly wide: any statement about how this system would behave on a new deployment carries much more uncertainty than the paired contrasts suggest.

We developed and evaluated this method at a single site. The prompting framework is sensor-agnostic, but the visual features learned from ground-camera RGB would not transfer to satellite thermal infrared without retraining, and site-specific factors (traffic density, wind climatology, camera geometry) may shift the optimal prompt design away from the 5-minute window found here.

This paper does not include a head-to-head comparison with detect-then-attribute pipelines such as Mask2Former (Cheng et al., 2021b, a) followed by post-hoc attribution (Dalmau et al., 2025), and we regard this as the most significant gap in the evaluation. A fair comparison requires a shared protocol controlling for detection recall, tracker fragmentation, and assignment ambiguity, each of which the two architectures handle in structurally different ways: a detection-based system segments all contrail pixels regardless of origin and assigns flight identity afterward, whereas ours poses a per-flight question and never produces an unattributed mask. Comparing segmentation quality alone would ignore the attribution and tracking dimensions that motivate the design, so the comparison has to run the full sequential pipeline on the same test set under the same metrics. We plan to address this in follow-up work. The structural argument, that intrinsic attribution eliminates a class of compounding errors, is a necessary part of the motivation but not a substitute for an empirical test.

7.0.1 Architectural modifications evaluated jointly

The five SAM2 modifications are applied as a package. Leave-one-out ablations are absent, so the claim that each transfers to other physics-prompted settings rests on mechanistic reasoning rather than direct evidence, and we cannot say which of the five carries the weight.

7.0.2 The negative-signal axis is not fully crossed

We trained age-weighted prompts with and without the negative signal, but never binary prompts with it. The negative signal is therefore only ever observed on top of an age gradient, and its effect cannot be separated from the encoding it is layered onto. The same applies to the negative signal's own representation: we tested only the age-weighted form, not a binary present/absent variant.

8 Conclusions

Contrail detection, tracking, and attribution collapse into a single forward pass when the model is conditioned on flight trajectory data. SAM2, prompted with spatial masks derived from ADS-B positions and ERA5 wind fields through CoCiP, checks each candidate flight's predicted location and either confirms the contrail with a refined mask or rejects the prompt. Tracking and attribution follow from the architecture: each prompt carries a flight identifier across all frames, so identity never needs to be re-established.

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

Telling the model how old each part of the contrail is (age-weighted encoding) lifts attribution precision from 89% to 96%, the single largest improvement in the ablation; because the binary baseline also uses a shorter window, encoding and window move together in that particular comparison (Sect. 5). Telling it where competing flights are (the negative signal) buys attribution recall rather than precision. Keeping the age window short, 5 rather than 10 minutes, avoids the detection penalty caused by wind-drifted prompts. Encoding richness and age window act largely independently within the grid, and the best combination reaches $\text{mAP}_{25:75}$ 0.380 at 96.2% attribution precision (Table 2).

It is worth being precise about what those numbers do and do not show. Richer prompts do not find more contrails: the detection rate and track completeness of the best variant are statistically indistinguishable from the binary baseline under video-level resampling. Nor do they draw better masks once a contrail is found, at least beyond the step from binary to age-weighted, since temporal IoU restricted to detected flights is flat across the whole 2×2 grid. What they change is which flight each mask is assigned to, and how much of a contrail's lifetime the model holds on to. For a monitoring application that is the useful half, because an unattributed contrail cannot validate anything.

The SAM2 modifications for dense prompting (memory attention on conditioning frames, frame-distance temporal encoding, soft memory sigmoid, dice-weighted loss, and routing every prompt through the mask decoder) address problems that arise whenever per-frame external prompts replace sparse human annotations. The modifications are not specific to contrails and are expected to transfer to other domains where a non-visual sensor provides per-frame spatial priors of similar density and structure, though retraining on domain-specific data will be required in each case. The prompt itself remains open to enrichment: local relative humidity over ice, persistence probability, and wind shear magnitude are all available from the same ERA5 and CoCiP pipeline, require no additional labeling, and could be embedded as extra prompt channels without any architectural change.

Physical constraints should inform visual recognition before the fact, not validate it afterward. A contrail cannot exist without a flight; encoding that constraint as a spatial prompt replaces a three-stage pipeline with a single model, and the same logic applies wherever external data constrains where objects can appear.

Code and data availability. The complete system, comprising training and inference code, the adapted SAM2 configurations, the evaluation suite implementing every metric in Sect. 5.1, and the scripts that generate all figures and statistical analyses in this paper, is open source at https://github.com/eurocontrol-asu/sam2-contrails, with pretrained checkpoints for all five prompt variants. The GVCCS dataset (Jarry et al., 2026) is publicly available.

Appendix A: Per-flight case studies

The figures in this appendix are single-flight illustrations of effects that Sects. 5 and 5.4 establish quantitatively across the test set. They are included because the mechanisms behind the prompt-design results are easier to see on a worked case than to infer from aggregate metrics, but no claim in the paper rests on them.

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

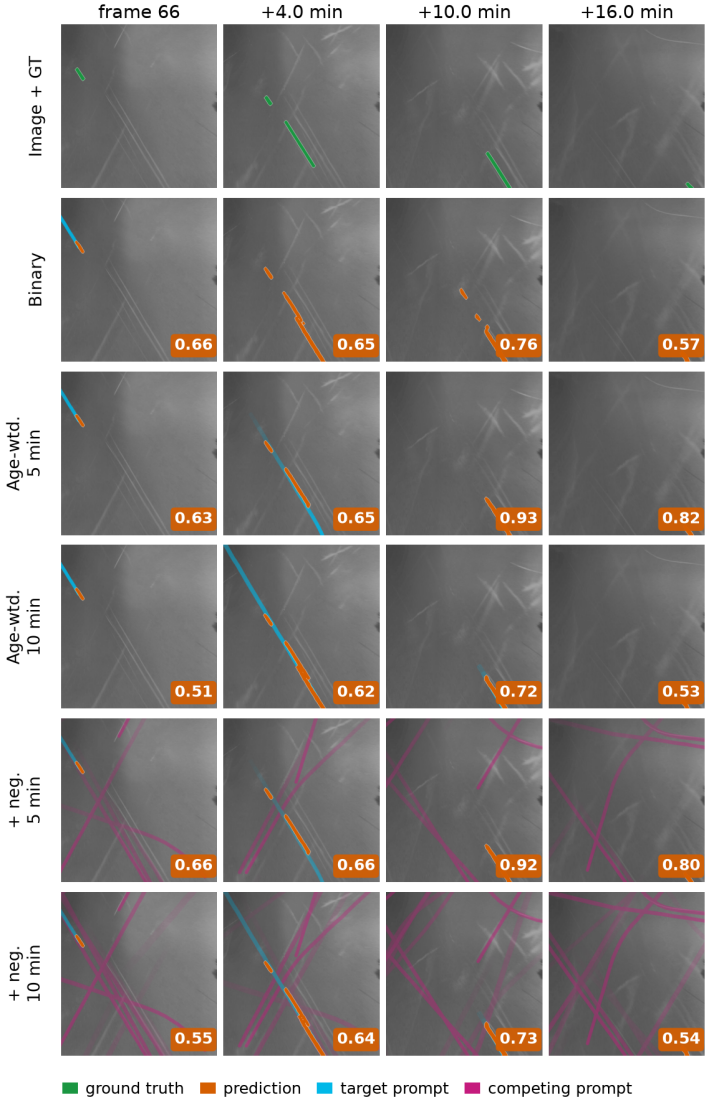

Figure A1. Flight 324b5d6a (video 12) tracked by all five variants; columns are labelled by elapsed time (30-second cadence). Row 1: image with ground truth overlay (green, thickened for visibility). Rows 2–6: per-variant prompt and prediction, using flat *cyan* for the binary prompt, *cyan* gradient for age-weighted prompts, plus *magenta* competing-flight suppression for the negative-signal variants; vermilion shows the predicted mask (thickened for visibility) with its object-score badge. All variants fire confidently on most frames, but only the 5-minute age-weighted variants keep the mask on the annotated contrail (completeness 89% without and 92% with the negative signal, against 31% for binary and 39%/33% for the drift-affected 10-minute variants). The last column falls after the annotation ends, so every score badge there is a prediction with no ground truth left to match.

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

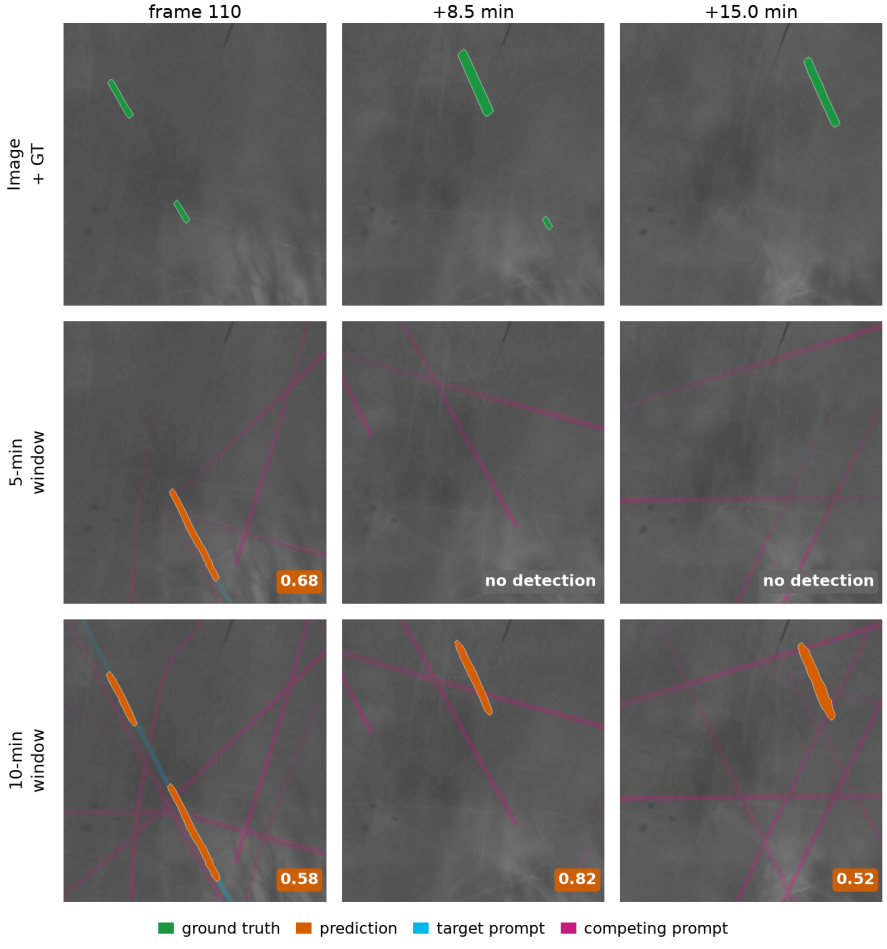

Figure A2. Age-window trade-off, case A: a late-visible contrail (flight 3363798b, video 5); columns are labelled by elapsed time (30-second cadence). Row 1: image with ground truth (green). Rows 2–3: 5- and 10-minute windows, with *cyan* target prompt, *magenta* competing-flight suppression, and vermilion predicted mask with score badge; frames without a scored mask are marked "no detection". The contrail becomes clearly visible only several minutes after formation: the 5-min window covers too small a footprint and its two scored predictions fail the IoU ≥ 0.25 match, while the 10-min window extends the prompt along the full contrail (2.7% vs. 70.3% completeness).

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

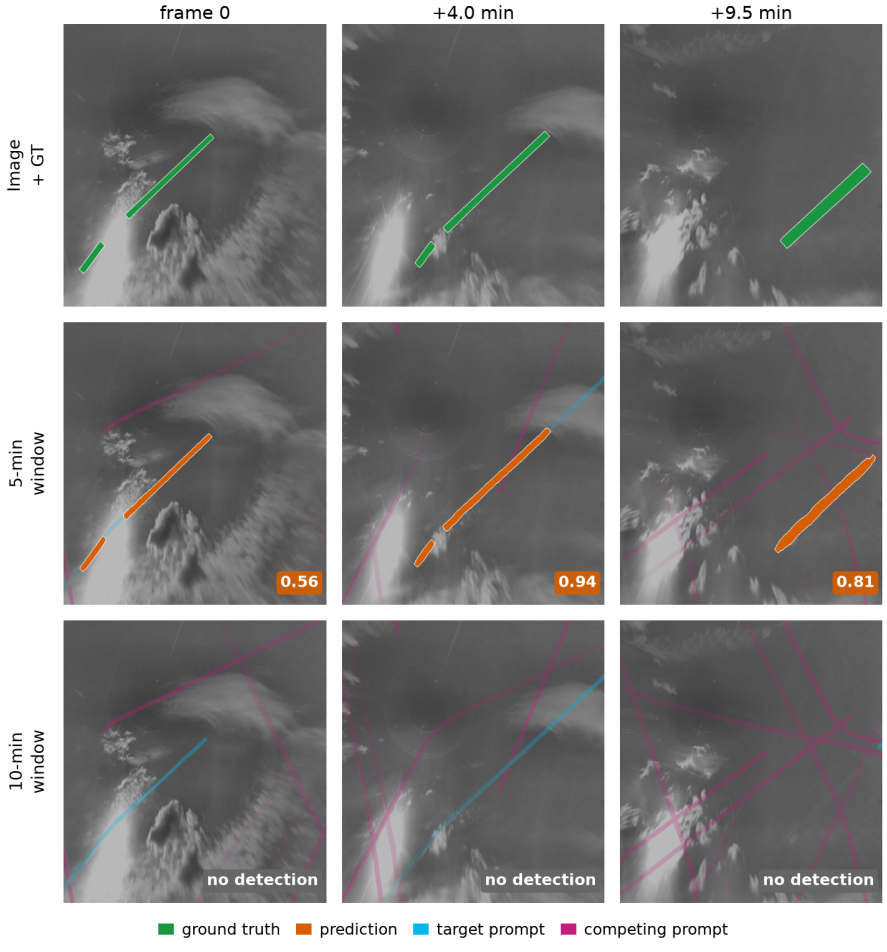

Figure A3. Age-window trade-off, case B: a crowded scene (flight 33c16444, video 7); same layout and colour coding as Fig. A2. Older unprompted contrails cross the scene: under the 10-min window they enter the negative signal and suppress the true detection on every frame (0% completeness), whereas the 5-min window avoids the contamination (100%).

Author contributions. RD: Formal analysis, Investigation, Methodology, Software, Visualization, Writing (original draft preparation). GJ: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing (original draft preparation). PV: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Resources, Software, Writing (original draft preparation).

Competing interests. The contact author has declared that none of the authors has any competing interests.

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

Acknowledgements. This work uses ERA5 reanalysis data from the Copernicus Climate Change Service, flight surveillance data collected through the OpenSky Network, and the `pycontrails` implementation of CoCiP. The authors thank the maintainers of these open resources, as well as the annotation effort behind the GVCCS dataset.

During the preparation of this work the authors used a large language model to assist with copy-editing and with checking the internal consistency of reported figures. The authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

References

Appleman, H.: The formation of exhaust condensation trails by jet aircraft, Bulletin of the American Meteorological Society, 34, 14–20, https://doi.org/10.1175/1520-0477-34.1.14, 1953.

Cheng, B., Choudhuri, A., Misra, I., Kirillov, A., Girdhar, R., and Schwing, A. G.: Mask2Former for video instance segmentation, arXiv preprint arXiv:2112.10764, https://doi.org/10.48550/arXiv.2112.10764, 2021a.

Cheng, B., Schwing, A., and Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation, in: Advances in Neural Information Processing Systems (NeurIPS), https://doi.org/10.48550/arXiv.2107.06278, 2021b.

Cheng, H. K. and Schwing, A. G.: XMem: Long-term video object segmentation with an Atkinson-Shiffrin memory model, in: European Conference on Computer Vision (ECCV), https://doi.org/10.1007/978-3-031-19815-1_37, 2022.

Cheng, H. K., Oh, S., Price, B., Schwing, A., and Lee, J.-Y.: Putting the object back into video object segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), https://doi.org/10.1109/CVPR52733.2024.00304, 2024.

Chevallier, R. et al.: Linear contrail detection, tracking and matching with aircraft using geostationary satellite and air traffic data, Remote Sensing, 15, https://doi.org/10.3390/rs15030740, 2023.

Dalmau, R., Jarry, G., and Very, P.: Contrail-to-flight attribution using ground visible cameras and flight surveillance data, in: 15th SESAR Innovation Days (SID 2025), Lake Bled, Slovenia, https://doi.org/10.48550/arXiv.2510.16891, paper 2025-024, preprint arXiv:2510.16891, 2025.

Geraedts, S. et al.: A scalable system to measure contrail formation on a per-flight basis, Environmental Research Communications, 6, 015 008, https://doi.org/10.1088/2515-7620/ad11ab, 2024.

Gourgue, N. et al.: A dataset of annotated ground-based images for the development of contrail detection algorithms, Data in Brief, 59, 111 354, https://doi.org/10.1016/j.dib.2025.111354, 2025.

He, K., Gkioxari, G., Dollár, P., and Girshick, R.: Mask R-CNN, in: IEEE International Conference on Computer Vision (ICCV), https://doi.org/10.1109/ICCV.2017.322, 2017.

Hersbach, H. et al.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Jarry, G., Dalmau, R., Very, P., Ballerini, F., and Bocu, S.-D.: GVCCS: a dataset for contrail identification and tracking on visible whole sky camera sequences, Earth System Science Data, 18, 1037–1059, https://doi.org/10.5194/essd-18-1037-2026, 2026.

Kirillov, A. et al.: Segment Anything, in: IEEE International Conference on Computer Vision (ICCV), https://doi.org/10.1109/ICCV51070.2023.00371, 2023.

Lee, D. S. et al.: The contribution of global aviation to anthropogenic climate forcing for 2000 to 2018, Atmospheric Environment, 244, 117 834, https://doi.org/10.1016/j.atmosenv.2020.117834, 2021.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P.: Focal loss for dense object detection, in: IEEE International Conference on Computer Vision (ICCV), https://doi.org/10.1109/ICCV.2017.324, 2017.

Lin, T.-Y. et al.: Microsoft COCO: common objects in context, in: European Conference on Computer Vision (ECCV), https://doi.org/10.1007/978-3-319-10602-1_48, 2014.

Loshchilov, I. and Hutter, F.: Decoupled weight decay regularization, in: International Conference on Learning Representations (ICLR), https://doi.org/10.48550/arXiv.1711.05101, 2019.

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

Low, J., Teoh, R., Ponsonby, J., Gryspeerdt, E., Shapiro, M., and Stettler, M. E. J.: Ground-based contrail observations: comparisons with reanalysis weather data and contrail model simulations, Atmospheric Measurement Techniques, 18, 37–56, https://doi.org/10.5194/amt-18-37-2025, 2025.

Mannstein, H., Meyer, R., and Wendling, P.: Operational detection of contrails from NOAA-AVHRR data, International Journal of Remote Sensing, 20, 1641–1660, https://doi.org/10.1080/014311699212650, 1999.

Ng, J. Y.-H. et al.: OpenContrails: Benchmarking contrail detection on GOES-16 ABI, arXiv preprint arXiv:2304.02122, https://doi.org/10.48550/arXiv.2304.02122, 2023.

Poll, D. I. A. and Schumann, U.: An estimation method for the fuel burn and other performance characteristics of civil transport aircraft in the cruise. Part 1: fundamental quantities and governing relations for a general atmosphere, The Aeronautical Journal, 125, 257–295, https://doi.org/10.1017/aer.2020.62, 2021.

Ravi, N. et al.: SAM 2: Segment Anything in Images and Videos, arXiv preprint arXiv:2408.00714, https://doi.org/10.48550/arXiv.2408.00714, 2024.

Riggi-Carrolo, E., Dubot, T., Sarrat, C., and Bedouet, J.: AI-driven identification of contrail sources: integrating satellite observation and air traffic data, Journal of Open Aviation Science, 1, https://doi.org/10.59490/joas.2023.7235, 2023.

Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), https://doi.org/10.1007/978-3-319-24574-4_28, 2015.

Ryali, C. et al.: Hiera: a hierarchical vision transformer without the bells-and-whistles, in: International Conference on Machine Learning (ICML), https://doi.org/10.48550/arXiv.2306.00989, 2023.

Sankar, T. et al.: Efficacy of scalable airline-led contrail avoidance, arXiv preprint arXiv:2603.06909, https://doi.org/10.48550/arXiv.2603.06909, 2026.

Sarna, A. et al.: Benchmarking and improving algorithms for attributing satellite-observed contrails to flights, Atmospheric Measurement Techniques, 18, 3495–3532, https://doi.org/10.5194/amt-18-3495-2025, 2025.

Schäfer, M., Strohmeier, M., Lenders, V., Martinovic, I., and Wilhelm, M.: Bringing up OpenSky: a large-scale ADS-B sensor network for research, in: Proceedings of the 13th IEEE/ACM International Symposium on Information Processing in Sensor Networks (IPSN), pp. 83–94, https://doi.org/10.1109/IPSN.2014.6846743, 2014.

Schmidt, E.: Die Entstehung von Eisnebel aus den Auspuffgasen von Flugmotoren, Schriften der Deutschen Akademie der Luftfahrtforschung, 44, 1–15, 1941.

Schumann, U.: On conditions for contrail formation from aircraft exhausts, Meteorologische Zeitschrift, 5, 4–23, https://doi.org/10.1127/metz/5/1996/4, 1996.

Schumann, U.: A contrail cirrus prediction model, Geoscientific Model Development, 5, 543–580, https://doi.org/10.5194/gmd-5-543-2012, 2012.

Shapiro, M., Engberg, Z., Teoh, R., Stettler, M. E. J., et al.: pycontrails: Python library for modeling aviation climate impacts, Zenodo, https://py.contrails.org, https://doi.org/10.5281/zenodo.7776686, 2024.

Sonabend-W, A. et al.: Feasibility test of per-flight contrail avoidance in commercial aviation, Communications Engineering, 3, 184, https://doi.org/10.1038/s44172-024-00329-7, 2024.

Syed Ariff, S. H. et al.: Evaluating SAM2 for video semantic segmentation, Machine Intelligence Research, https://doi.org/10.1007/s11633-026-1638-9, arXiv:2512.01774, 2025.

https://doi.org/10.5194/egusphere-2026-4608
Preprint. Discussion started: 12 August 2026

Teoh, R. et al.: Mitigating the climate forcing of aircraft contrails by small-scale diversions and technology adoption, Environmental Science & Technology, 54, 2941–2950, https://doi.org/10.1021/acs.est.9b05608, 2020.

Teoh, R. et al.: Global aviation contrail climate effects from 2019 to 2021, Atmospheric Chemistry and Physics, 23, 11 303–11 319, https://doi.org/10.5194/acp-23-11303-2023, 2023.

Van Huffel, J., Ehrmanntraut, R., and Croes, D.: Contrail detection and classification using computer vision with ground-based cameras, in: 2025 Integrated Communications, Navigation and Surveillance Conference (ICNS), pp. 1–6, IEEE, https://doi.org/10.1109/ICNS65417.2025.10976944, 2025.

Xu, J. et al.: SAM4D: segment anything in camera and LiDAR streams, in: IEEE/CVF International Conference on Computer Vision (ICCV), https://doi.org/10.1109/ICCV51701.2025.02650, arXiv:2506.21547, 2025.

Yang, Z., Wei, Q., and Yang, Y.: Decoupling features in hierarchical propagation for video object segmentation, in: Advances in Neural Information Processing Systems (NeurIPS), https://doi.org/10.52202/068431-2632, 2022.

Zhang, J. and Tang, H.: SAM2 for image and video segmentation: a comprehensive survey, arXiv preprint arXiv:2503.12781, https://doi.org/10.48550/arXiv.2503.12781, 2025.

Zhou, Y. et al.: When SAM2 meets video camouflaged object segmentation: a comprehensive evaluation and adaptation, Visual Intelligence, https://doi.org/10.1007/s44267-025-00082-1, arXiv:2409.18653, 2025.