ASYNCHRONOUS MULTIMODAL CANDIDATE REWEIGHTING FOR MULTI-HORIZON PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-horizon prediction of real-world systems often requires reasoning over multiple candidate futures using heterogeneous observations acquired at different times. These candidates may be produced by physical simulators, learned predictors, or ensembles, yet existing fusion methods commonly assign each candidate a single reliability score that is reused across all prediction horizons. This assumption becomes restrictive as candidate futures diverge over time and the relevance of asynchronous evidence changes with the prediction target. We introduce Asynchronous Multimodal Candidate Reweighting (AMCR), a general framework that estimates reliability jointly over candidate identity and target time. Rather than synchronizing observations to a common clock, AMCR preserves their native timestamps together with modality, availability, and quality, and constructs target-conditioned contexts through modality-aware temporal relevance weighting. The resulting candidate-time reliability matrix enables the same candidate bank to be reweighted differently across prediction horizons. We establish a bounded-influence result showing that the effect of temporally distant evidence decays under the proposed temporal gate, together with a sufficient condition for candidate-selection stability. We evaluate AMCR under controlled timestamp perturbations, missing modalities, evidence-quality variation, and horizon-dependent candidate changes, and further instantiate it for physics-based oil-spill evolution using satellite imagery, vessel trajectories, environmental forcing, and simulator-generated futures.

1 INTRODUCTION

Multi-horizon prediction is often supported by heterogeneous observations that arrive asynchronously. Images may be sparse, trajectories irregular, behavior cues interval-valued, and environmental products updated on independent schedules. Conventional multimodal fusion can obscure this temporal structure by resampling observations to a common clock or collapsing them into one scene-level representation (Guo et al., 2024; Zhang et al., 2024). At the same time, many forecasting systems retain multiple candidate futures rather than a single deterministic prediction. A static candidate weight w_k is then restrictive because both candidate quality and evidence relevance can change with the requested horizon.

Oil-spill forecasting provides a representative instance of this setting. Physics-based models such as OpenDrift/OpenOil, MEDSLIK-II, GNOME, and Parcels can generate alternative spill evolutions under uncertain forcing and configuration (Dagestad et al., 2018; De Dominicis et al., 2013; Beegle-Krause, 2001; Delandmeter & van Sebille, 2019). These candidates are evaluated using external evidence that is not synchronized with the simulator: SAR/optical observations provide sparse spatial measurements, AIS contributes irregular vessel trajectories and behavior context, and environmental fields provide time-varying physical constraints. The forecasting problem is therefore not to make simulation depend on all modalities, but to determine how the available evidence should change the reliability of each persistent future at each horizon.

We introduce AMCR, a candidate-preserving framework that treats multi-horizon forecasting as evidence-guided reliability reasoning over persistent futures. Native-time observations are retained as target-conditioned evidence tokens, while each candidate is represented through a horizon-specific fusion embedding derived from its evolving state. AMCR then relates these two representations through temporal, spatial, activity, and physical consistency, forming relation-evidence embeddings that

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

drive signed evidence propagation over candidate–time nodes. The resulting scores are normalized independently at each horizon, allowing the same candidate bank to be reweighted differently as both the candidate states and the relevance of asynchronous observations change. Candidate construction is treated as an upstream stage, so the reweighting mechanism can operate over simulator-derived, learned, or ensemble futures.

AMCR centers reliability estimation on the candidate-time relation rather than on a single global candidate score. Evidence is interpreted relative to a specific candidate and target horizon, enabling the model to distinguish supportive, contradictory, and unresolved observations instead of collapsing them into a single confidence signal. This relation-aware formulation is coupled with native-time temporal gating and horizon-wise normalization, providing interpretable reliability values $w_{k,h}$ while preserving causal availability of evidence. The framework further admits a bounded-influence property for temporally distant observations and a sufficient score-margin condition for stability under bounded perturbations. Together, these elements provide a general mechanism for asynchronous multimodal reweighting in which candidate preference can evolve across horizons without discarding the persistent structure of the underlying future hypotheses.

2 Related Work

Multimodal reasoning, uncertainty, and graph interaction. Modern remote-sensing encoders increasingly exploit heterogeneous sensors and multimodal context (Guo et al., 2024; Zhang et al., 2024). AIS representation learning provides additional vessel-activity context for maritime reasoning (Nguyen et al., 2022). Graph attention provides a natural mechanism for relational reasoning among structured hypotheses (Veličković et al., 2018), while uncertainty and calibration methods emphasize that confidence should be evaluated separately from point accuracy (Lakshminarayanan et al., 2017; Guo et al., 2017). AMCR brings these directions together for a specific inference problem: reweighting a persistent bank of oil-spill candidate futures as evidence relevance and candidate quality change with horizon.

Oil-spill observation and segmentation. Satellite remote sensing, especially SAR, is central to marine oil-spill detection because slicks can alter surface roughness and radar backscatter (Brekke & Solberg, 2005; Fingas & Brown, 2014; Skrunes et al., 2014; Al-Ruzouq et al., 2020). Deep models have improved segmentation and discrimination of spill-like structures in SAR and optical imagery (Krestenitis et al., 2019; Bianchi et al., 2020; Chen et al., 2023; Fan et al., 2024). These methods estimate the observed spill state but do not directly address which future simulation should be trusted after the detected anchor.

Physics-based drift forecasting. Lagrangian systems such as OpenDrift/OpenOil, MEDSLIK-II, GNOME, and Parcels model transport under environmental forcing (Dagestad et al., 2018; De Dominicis et al., 2013; Beegle-Krause, 2001; Delandmeter & van Sebille, 2019). Their forecasts depend on forcing quality and configuration; recent work shows that correcting forecast sea-surface dynamics can improve downstream oil-drift prediction (Ren et al., 2025). AMCR is complementary: rather than replacing the simulator, it retains multiple physical hypotheses and learns horizon-specific reliability over them.

3 Problem Formulation

3.1 Asynchronous Multimodal Multi-Horizon Setting

We consider a multi-horizon prediction problem with two inputs: a persistent bank of candidate futures and heterogeneous evidence observed on different temporal schedules. A *candidate* describes a possible future, whereas an *evidence item* provides observed information that may support, contradict, or leave unresolved a candidate at a target horizon.

Let

$$\mathcal{C} = \{C_k\}_{k=1}^{K}, \qquad C_k = \{Y_{k,h}\}_{h=1}^{H}, \qquad \mathcal{T} = \{t_1, \ldots, t_H\}, \tag{1}$$

where C_k is a persistent future hypothesis and $Y_{k,h}$ its state at target t_h. Candidates may come from physical simulators, learned predictors, or ensembles; AMCR only requires candidate identity to be preserved across horizons.

The multimodal evidence set is

$$\mathcal{E} = \{e_i\}_{i=1}^{N}, \qquad e_i = (x_i, m_i, \tau_i^{\mathrm{occ}}, \tau_i^{\mathrm{obs}}, \Omega_i, q_i, u_i, \xi_i, a_i), \tag{2}$$

where x_i is observed content, m_i the modality, τ_i^{occ} the physical occurrence or valid time, and τ_i^{obs} the time at which the item becomes available to the forecasting system. The remaining terms encode spatial/contextual support, quality, uncertainty, provenance, and availability.

This distinction is central to asynchronous multimodal reasoning: an event can occur before it becomes observable, and different modalities have different temporal support and sampling schedules. Satellite imagery provides sparse snapshots, AIS provides irregular trajectory observations, behavior cues may span intervals, and environmental products follow their own valid/update times. Hence multimodality involves both heterogeneous content and heterogeneous temporal structure.

For target t_h, causal availability and temporal relevance are separated as

$$a_{i,h} = a_i \, \mathbb{I}[\tau_i^{\mathrm{obs}} \le t_h], \qquad g_{i,h} = a_{i,h} \exp[-d_{m_i}(\tau_i^{\mathrm{occ}}, t_h)], \qquad d_m \ge 0, \tag{3}$$

where d_{m_i} is a modality-aware temporal distance. Thus, evidence cannot influence a target before it becomes available, while its relevance is determined by when the underlying event or state occurred. AMCR therefore preserves native temporal information rather than synchronizing all modalities to a common clock.

Even with a fixed candidate bank, the evidence available and relevant to t_h can differ from that at t_{h+1}. Candidate reliability must therefore depend jointly on candidate identity and target time rather than being treated as a fixed property.

Oil-spill instantiation. Oil-spill forecasting directly exhibits this setting. A physics-based candidate bank provides alternative evolutions over 24, 48, 72 h and other horizons, while SAR/optical imagery, AIS trajectories and behavior, and environmental fields provide asynchronous external evidence. Their occurrence, acquisition, and availability times need not coincide. These observations evaluate persistent candidate futures rather than becoming part of the simulator itself, while $Y_{k,h}$ contains the candidate spill field and associated geometric, drift, and uncertainty descriptors.

3.2 Candidate–Time Reweighting Objective

Candidate quality need not remain stationary across the forecast window. If $Y_h^\star$ denotes an available reference and $d(\cdot,\cdot)$ a task discrepancy, the horizon-wise oracle is

$$k_h^\star = \arg\min_k d(Y_{k,h}, Y_h^\star). \tag{4}$$

There is no requirement that $k_h^\star$ remain constant with h. A horizon-invariant weight w_k is therefore a restricted special case of the candidate–time reliability matrix

$$W = [w_{k,h}], \qquad w_{k,h} = p(k \mid \mathcal{E}_h^{\mathrm{avail}}, t_h), \qquad w_{k,h} \ge 0, \qquad \sum_{k=1}^{K} w_{k,h} = 1, \tag{5}$$

where

$$\mathcal{E}_h^{\mathrm{avail}} = \{e_i \in \mathcal{E} : a_i = 1,\ \tau_i^{\mathrm{obs}} \le t_h\} \tag{6}$$

denotes the evidence causally available for target t_h. Its effective contribution is further modulated by the native-time relevance $g_{i,h}$ in Eq. 3.

The generic fused prediction is

$$\widehat{Y}_h = \sum_{k=1}^{K} w_{k,h} Y_{k,h}. \tag{7}$$

This formulation is task-agnostic: $Y_{k,h}$ can represent any prediction object that can be combined across candidates. In the oil-spill instantiation, each candidate state is represented as a spatial

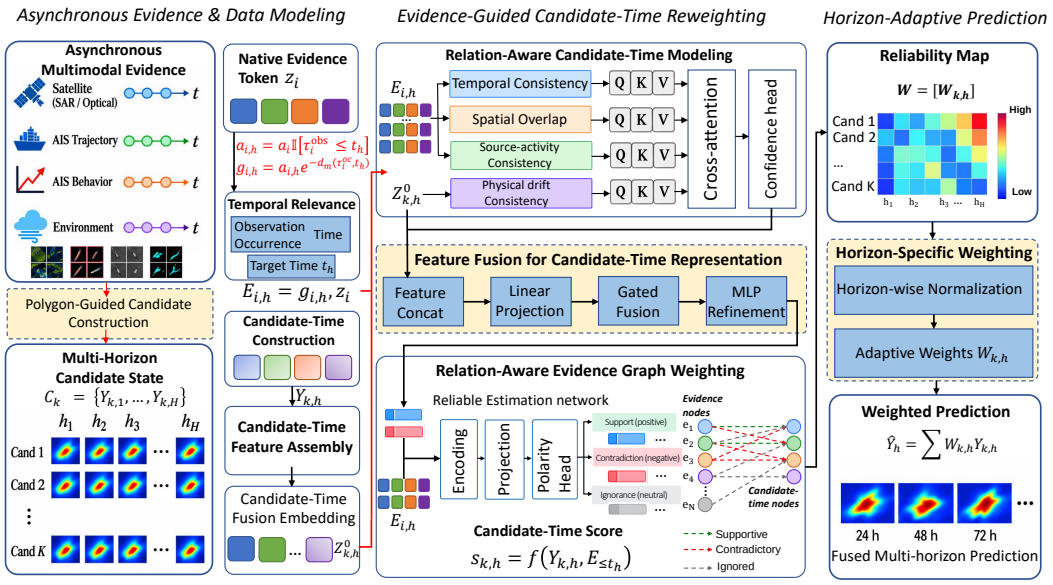

Figure 1: AMCR overview. Target-conditioned evidence tokens $E_{i,h}$ and candidate-time embeddings $z^0_{k,h}$ are coupled through relation-aware modeling and signed evidence weighting to produce candidate-time scores $s_{k,h}$. Horizon-wise normalization yields reliability weights $W = [w_{k,h}]$ and the fused multi-horizon prediction.

spill-probability field $P_k^{\text{spill}}(p, t_h)$, so the generic fusion specializes pointwise over spatial location p:

$$P^{\text{evo}}(p, t_h) = \sum_{k=1}^{K} w_{k,h} P_k^{\text{spill}}(p, t_h). \tag{8}$$

AMCR therefore does not seek a single candidate that is globally reliable across the complete forecast window. Instead, it estimates how much each persistent candidate should contribute at each horizon under evidence that is both *causally available* and *temporally relevant*. As new evidence becomes available or existing observations become less relevant to later targets, reliability is redistributed over the same candidate bank, motivating reweighting at the candidate–time level.

Reweighting cannot recover futures absent from $\mathcal{C}$, so candidate-bank coverage and evidence-guided reasoning are evaluated separately. An oracle measures the coverage of the fixed candidate bank, while AMCR learns the causal mapping

$$(\mathcal{C}, \mathcal{E}, \mathcal{T}) \mapsto W \tag{9}$$

that adapts candidate preference to changes in evidence availability, temporal relevance, and candidate evolution across horizons.

4 Proposed Method: Asynchronous Multimodal Candidate Reweighting

AMCR consists of three modules. *Asynchronous Evidence & Data Modeling* preserves native occurrence and observation times while representing the persistent candidate bank. *Evidence-Guided Candidate–Time Reweighting* couples evidence and candidate-time embeddings through relation-aware signed reasoning, and *Horizon-Adaptive Prediction* produces horizon-specific candidate weights and fused predictions, as summarized in Fig. 1.

4.1 Asynchronous Evidence & Data Modeling

The first module constructs target-conditioned evidence tokens and candidate-time embeddings. For oil-spill forecasting, the observed spill polygon guides the persistent candidate bank, while asynchronous observations remain external evidence for reweighting.

Native-time multimodal evidence. For evidence item e_i, let x_i denote content, m_i modality, τ_i^{occ} its physical occurrence or valid time, τ_i^{obs} the time it becomes observable, Ω_i spatial/contextual support, q_i quality, u_i uncertainty, ξ_i provenance, and a_i availability. A modality-specific encoder produces

$$z_i = \phi_{m_i}(x_i) + E_m(m_i) + E_{\mathrm{occ}}(\tau_i^{\mathrm{occ}} - T_0) + E_{\mathrm{obs}}(\tau_i^{\mathrm{obs}} - T_0) + E_\Omega(\Omega_i) + E_q(q_i) + E_u(u_i) + E_\xi(\xi_i). \tag{10}$$

Intrinsic reliability is

$$R_i^{\mathrm{node}} = \sigma\big(a_q q_i - a_u \tilde{u}_i + w_\xi^\top E_\xi(\xi_i) + b_R\big), \qquad a_q, a_u > 0, \tag{11}$$

where $\tilde{u}_i \in [0,1]$ is normalized uncertainty. For target t_h,

$$a_{i,h} = a_i \mathbb{I}[\tau_i^{\mathrm{obs}} \le t_h], \qquad g_{i,h} = a_{i,h} \exp[-d_{m_i}(\tau_i^{\mathrm{occ}}, t_h)], \qquad E_{i,h} = g_{i,h} z_i, \tag{12}$$

with $d_m(\tau_i^{\mathrm{occ}}, t_h) = |\tau_i^{\mathrm{occ}} - t_h|/\sigma_m$. Thus availability depends on when evidence becomes observable, whereas temporal relevance depends on the occurrence-to-target distance. This preserves the distinct schedules of satellite, AIS, behavior, and environmental evidence.

Polygon-guided candidate construction. For oil-spill forecasting, the detected geometry at T_0 provides an anchor polygon for constructing $\mathcal{C} = \{C_k\}_{k=1}^K$. Each candidate preserves its identity across horizons, while AIS, environmental, and other asynchronous observations remain external evidence for AMCR. Details are given in Appendix A.1.

Multi-horizon candidate states. A persistent candidate $C_k = \{Y_{k,h}\}_{h=1}^H$ is represented using its current state, temporal evolution, and cross-candidate context:

$$\begin{aligned} \nu_{k,h}^{\mathrm{self}} &= \psi_s(Y_{k,h}, Y_{k,h-1}, Y_0), \\ \nu_{k,h}^{\mathrm{cross}} &= \frac{1}{K-1} \sum_{j \neq k} \psi_c(Y_{k,h}, Y_{j,h}), \\ q_{k,h} &= [\phi_C(C_k); \phi_Y(Y_{k,h}); \phi_\nu(\nu_{k,h}^{\mathrm{self}}, \nu_{k,h}^{\mathrm{cross}}); \phi_t(t_h - T_0)]. \end{aligned} \tag{13}$$

For the maritime case, ϕ_ν includes polygon deformation, centroid displacement, drift direction/speed, fragmentation, uncertainty, and candidate agreement. The candidate-time embedding is

$$z_{k,h}^0 = F_{\mathrm{CT}}(q_{k,h}), \tag{14}$$

where F_{CT} is a shared projection/fusion network.

4.2 Evidence-Guided Candidate–Time Reweighting

The second module evaluates how each available evidence item affects candidate k at horizon h. Since evidence availability and temporal relevance vary with target time, the same candidate can be assessed under different evidence contexts. AMCR therefore relates $E_{i,h}$ directly to $z_{k,h}^0$ rather than assigning a global evidence weight.

Relation-aware candidate-time modeling. For each admissible pair (i,k,h),

$$\delta_{i,k,h}^{(f)} = D_f(E_{i,h}, z_{k,h}^0), \qquad f \in \mathcal{F} = \{\mathrm{temporal}, \mathrm{spatial}, \mathrm{activity}, \mathrm{physical}\}. \tag{15}$$

The factors capture temporal alignment, spatial overlap, source/activity consistency, and physical-drift consistency. Temporal alignment uses the occurrence-to-target relation, while causal availability is determined by τ_i^{obs}. Each factor produces

$$r_{i,k,h}^{(f)} = \phi_f([E_{i,h}; z_{k,h}^0; \delta_{i,k,h}^{(f)}]). \tag{16}$$

Cross-relation attention gives

$$\alpha_{i,k,h}^{(f)} = \mathrm{softmax}_f\left(\frac{(Q_f z_{k,h}^0)^\top K_f r_{i,k,h}^{(f)}}{\sqrt{d}}\right), \qquad \bar{r}_{i,k,h} = \sum_{f\in\mathcal{F}} \alpha_{i,k,h}^{(f)} V_f r_{i,k,h}^{(f)}. \tag{17}$$

A confidence head estimates

$$\gamma_{i,k,h} = \sigma\big(f_{\mathrm{conf}}([\bar{r}_{i,k,h}; E_{i,h}; z_{k,h}^0])\big), \tag{18}$$

so $(\bar{r}_{i,k,h}, \gamma_{i,k,h})$ represents how evidence i constrains candidate k at horizon h.

Feature fusion for candidate-time representation. The candidate-time, evidence, and relation features are fused as

$$p_{i,k,h} = W_P[z_{k,h}^0; E_{i,h}; \bar{r}_{i,k,h}; \gamma_{i,k,h}], \qquad G_{i,k,h} = \sigma(W_G p_{i,k,h} + b_G), \tag{19}$$

followed by

$$H_{i,k,h} = \mathrm{MLP}_F\big(\mathrm{LN}\big[W_Z z_{k,h}^0 + G_{i,k,h} \odot p_{i,k,h}\big]\big). \tag{20}$$

The resulting $H_{i,k,h}$ is passed to graph weighting while preserving the evidence index.

Relation-aware evidence graph weighting. The reliability estimator predicts evidence polarity:

$$u_{i,k,h} = f_{\mathrm{enc}}(H_{i,k,h}), \qquad \tilde{u}_{i,k,h} = f_{\mathrm{proj}}(u_{i,k,h}), \qquad [\pi_{i,k,h}^+, \pi_{i,k,h}^-, \pi_{i,k,h}^0] = \mathrm{softmax}(f_{\mathrm{pol}}(\tilde{u}_{i,k,h})), \tag{21}$$

where π^+, π^-, and π^0 denote support, contradiction, and ignorance. Unsigned trust is

$$T_{i,k,h} = g_{i,h} R_i^{\mathrm{node}} \gamma_{i,k,h}(1 - \pi_{i,k,h}^0), \tag{22}$$

which controls message magnitude independently of polarity. Since $g_{i,h} = 0$ before τ_i^{obs} and decays with occurrence-to-target distance afterward, unavailable or stale evidence is suppressed.

Evidence and candidate-time states form a bipartite graph with

$$\ell_{i,k,h} = \frac{(Q z_{k,h}^0)^\top K H_{i,k,h}}{\sqrt{d}} + \lambda_T \log(T_{i,k,h} + \epsilon), \qquad \beta_{i,k,h} = \mathrm{MaskedSoftmax}_i(\ell_{i,k,h}; a_{i,h}). \tag{23}$$

The signed message is

$$m_{i\to(k,h)} = \beta_{i,k,h}\left[\pi_{i,k,h}^+ W_+ H_{i,k,h} - \pi_{i,k,h}^- W_- H_{i,k,h}\right], \tag{24}$$

and aggregation gives

$$\tilde{z}_{k,h} = \mathrm{LN}\left(z_{k,h}^0 + \sum_i m_{i\to(k,h)}\right), \qquad s_{k,h} = f_{\mathrm{score}}(\tilde{z}_{k,h}). \tag{25}$$

Thus relation modeling determines relevance, polarity determines support or contradiction, and graph weighting controls evidential influence.

Bounded influence of temporally distant evidence. Let $\tilde{\ell}_{i,k,h}$ collect the non-trust terms in Eq. 23, assume $|\tilde{\ell}_{i,k,h} - \tilde{\ell}_{j,k,h}| \le \Delta$, and $T_{j,k,h} \ge \tau > 0$. Then

$$\frac{\beta_{i,k,h}}{\beta_{j,k,h}} \le \exp(\Delta)\left(\frac{T_{i,k,h}+\epsilon}{\tau+\epsilon}\right)^{\lambda_T} \le \exp(\Delta)\left(\frac{g_{i,h}+\epsilon}{\tau+\epsilon}\right)^{\lambda_T}. \tag{26}$$

Hence unavailable evidence has no effect and temporally distant evidence has bounded influence, while nearby reliable contradictions can retain negative impact.

4.3 HORIZON-ADAPTIVE PREDICTION

Candidate-time scores are normalized independently at each horizon:

$$w_{k,h} = \frac{\exp(s_{k,h})}{\sum_{j=1}^{K} \exp(s_{j,h})}, \qquad W = [w_{k,h}] \in \mathbb{R}^{K \times H}. \tag{27}$$

The same candidate can therefore receive different weights as candidate states, evidence availability, and temporal relevance evolve across horizons. The fused prediction is

$$\widehat{Y}_h = \sum_{k=1}^{K} w_{k,h} Y_{k,h}, \tag{28}$$

with $Y_{k,h} = P_k^{\mathrm{spill}}(\cdot, t_h)$ for oil-spill forecasting.

To limit unnecessary variation across similar horizons, candidate divergence and evidence-relevance change are

$$D_h = \frac{2}{K(K-1)} \sum_{j<k} d_y(Y_{j,h}, Y_{k,h}), \qquad \Delta_h^E = \frac{1}{N} \sum_i |g_{i,h} - g_{i,h-1}|, \tag{29}$$

with adaptive regularization

$$\lambda_h^{\mathrm{stb}} = \lambda_0 \exp[-\kappa(D_h + \zeta \Delta_h^E)], \qquad \mathcal{L}_{\mathrm{stb}} = \sum_{h=2}^{H} \lambda_h^{\mathrm{stb}} \, \mathrm{JSD}(w_{\cdot,h} \| w_{\cdot,h-1}). \tag{30}$$

Candidate-selection stability. Let $k_h^* = \arg\max_k s_{k,h}$ with margin

$$\Delta_h^{\mathrm{score}} = s_{k_h^*,h} - \max_{j \neq k_h^*} s_{j,h}. \tag{31}$$

If $|s'_{k,h} - s_{k,h}| \leq \delta_h$ for all candidates, then

$$\Delta_h^{\mathrm{score}} > 2\delta_h \quad \Longrightarrow \quad \arg\max_k s'_{k,h} = k_h^*. \tag{32}$$

Hence candidate selection remains unchanged under bounded perturbations when the score margin is sufficiently large.

Candidate generation is treated as an upstream, non-differentiable stage. Training combines fused-output supervision, soft per-horizon candidate-order supervision when available, adaptive stability regularization, and calibration. The final outputs are the horizon-indexed reliability matrix W and the fused prediction.

5 EXPERIMENTS AND EVALUATION

5.1 EXPERIMENTAL PROTOCOL

Dataset and candidate bank. We evaluate AMCR on satellite-based oil-spill observations from the Korean Exclusive Economic Zone (EEZ), combining KOMPSAT-5 SAR, Sentinel optical imagery (European Space Agency, 2026b), AIS context, environmental forcing, and simulator-derived candidates. Sentinel-1 SAR data (European Space Agency, 2026a) support the observation and polygon-seed stage. Each event contains an anchor scene at T_0, a spill-like mask/polygon, forcing and release assumptions, and a fixed model-visible bank of $K = 6$ simulator candidates selected from 12 deterministic generation configurations. Candidates from the same event remain in the same data split to prevent leakage. Candidate fields are cached before training so all methods use the same hypothesis bank, with evaluation at 24, 48, and 72 h.

Evaluation metrics. Spatial fidelity is measured using Mask Dice and IoU on Eq. (8), together with Hausdorff distance and centroid error, both reported in kilometers. Origin Hit@r measures whether the predicted source location lies within radius $r = 5$ km of the reference source location. Candidate ranking uses Top-1, Top-3, and Mean Reciprocal Rank (MRR), while Reject AP measures suppression of inconsistent hypotheses. Reliability is assessed using Expected Calibration Error (ECE), Brier score, and selective accuracy at 90% and 70% coverage.

Table 1: Oil-spill evolution, localization, and calibration. Dice/IoU evaluate fused masks; Hausdorff and centroid error evaluate localization; ECE measures calibration.

Model	Mask Dice↑	**Mask IoU↑**	Hausdorff ↓	**Centroid Err.↓**	Origin Hit@r↑	**ECE ↓**
Satellite mask only	0.4825	0.3214	34.28	6.12	–	0.3815
Single simulator	0.6312	0.4871	22.46	3.74	0.3125	0.2634
Uniform simulator ensemble	0.5874	0.4392	26.18	4.38	0.2750	0.3142
Uncertainty-weighted ensemble	0.7215	0.5849	18.27	2.71	0.4625	0.2074
Naive multimodal fusion	0.7896	0.6592	15.94	1.96	0.5375	0.1658
AMCR w/o polygon variation	0.8726	0.7813	12.78	0.91	0.6250	0.0679
AMCR-Static (w_k)	0.9021	0.8358	11.24	0.67	0.6750	0.0527
AMCR ($w_{k,h}$)	**0.9384**	**0.8872**	**9.76**	**0.38**	**0.7250**	**0.0367**

Table 2: Candidate ranking, rejection, and selective reliability. Higher is better except Brier score.

Model	Top-1 ↑	Top-3 ↑	MRR ↑	Reject AP↑	Acc@90 ↑	Acc@70 ↑	Brier ↓
Single simulator	0.5417	0.7083	0.6129	0.5836	0.5625	0.6875	0.3241
Uncertainty-weighted ensemble	0.6667	0.7917	0.7218	0.6815	0.6667	0.7708	0.2294
Naive multimodal fusion	0.7292	0.8750	0.7865	0.7421	0.7396	0.8333	0.1736
AMCR w/o contradiction channel	0.8333	0.9583	0.8889	0.8543	0.8438	0.9167	0.0876
AMCR	**0.8750**	**0.9583**	**0.9167**	**0.9128**	**0.9063**	**0.9583**	**0.0743**

Baselines and ablations. *Satellite mask only* propagates the anchor observation without explicit evolution, while *Single simulator* uses one OpenDrift/OpenOil hypothesis. *Uniform ensemble* averages candidates equally, and *uncertainty-weighted ensemble* uses simulator uncertainty for weighting. *Naive multimodal fusion* concatenates satellite, AIS, environmental, polygon, and candidate features without relation-aware or signed evidence reasoning. Ablations remove polygon variation, replace $w_{k,h}$ with horizon-invariant w_k (*AMCR-Static*), or remove the contradiction path in Eq. 24.

Implementation, optimization, and training details are provided in Appendix A.2.

5.2 Quantitative Results

Spatial evolution and calibration. Table 1 shows a consistent progression from single-model and non-learned fusion toward evidence-guided candidate-time reasoning. Uniform averaging underperforms the single simulator because equally mixing incompatible trajectories can blur the evolution field. Uncertainty-only weighting improves Mask IoU to 0.5849, while naive multimodal fusion reaches 0.6592. Explicit polygon variation improves deformation-aware comparison, and horizon-dependent reasoning yields the strongest spatial and calibration results: AMCR reaches Mask IoU 0.8872, centroid error 0.38, and ECE 0.0367, compared with 0.8358, 0.67, and 0.0527 for the horizon-invariant AMCR-Static control. Because both variants use the same candidate universe and evidence sources, this gap isolates the value of allowing reliability to vary with horizon rather than relying on a stronger upstream simulator.

Candidate reliability and selective prediction. Table 2 shows that AMCR also improves candidate-level reliability, obtaining Top-1 accuracy 0.8750, MRR 0.9167, Reject AP 0.9128, and Brier score 0.0743. Uncertainty-weighted fusion is weaker because simulator uncertainty is only one cue and cannot distinguish reliable support from reliable contradiction. Removing the contradiction channel lowers Reject AP to 0.8543 and worsens Brier to 0.0876, supporting the explicit separation of support, contradiction, and ignorance in Eqs. 21–24.

5.3 Qualitative Evolution Analysis

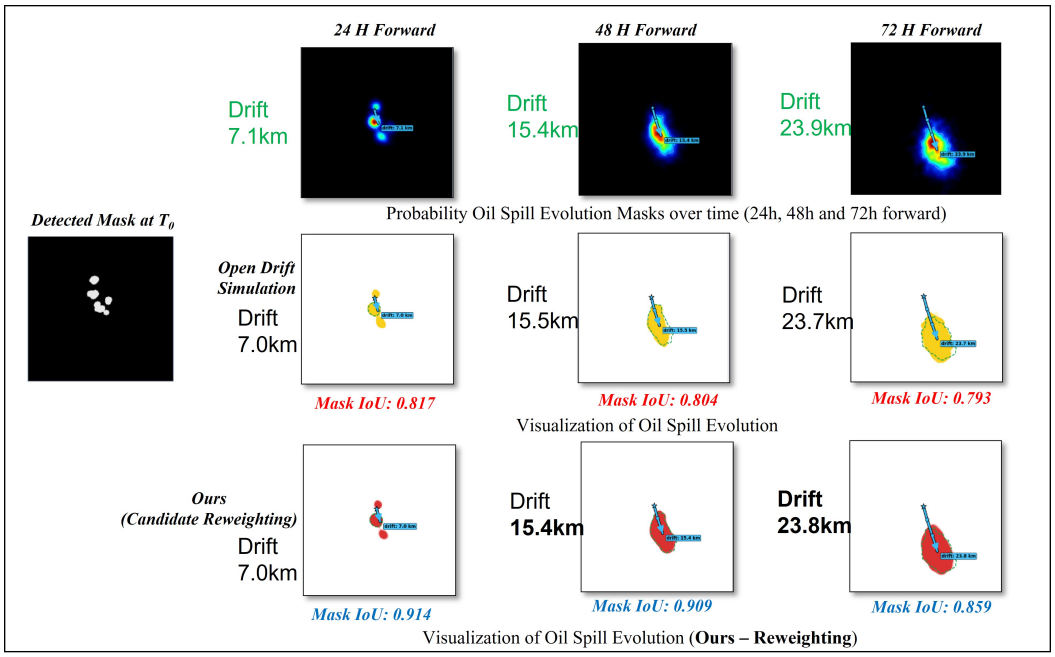

Figure 2: Forward evolution at 24, 48, and 72 h. AMCR reweights simulator candidates using the available multimodal evidence to better match the probabilistic reference spill geometry than a single OpenDrift trajectory.

Forward evolution. Figure 2 provides a direct qualitative view of the fused evolution. OpenDrift obtains Mask IoU $0.817/0.804/0.793$, while AMCR reaches $0.914/0.909/0.859$ at 24, 48, and 72 h, corresponding to gains of 0.097, 0.105, and 0.066. The candidate-reweighted masks remain closer to the probabilistic reference shapes as the footprint expands, whereas the drift-distance annotations remain comparable. The result illustrates the role of AMCR as an evidence-guided forecasting layer: persistent physical candidates are retained, but their contribution is adjusted separately at each horizon according to asynchronous satellite, AIS, behavior, and environmental evidence rather than treating one simulator realization as globally reliable.

6 Conclusion

We presented *Asynchronous Multimodal Candidate Reweighting* (AMCR), a candidate-preserving framework for multi-horizon prediction under asynchronous and heterogeneous evidence. AMCR models reliability at the candidate–time level, allowing candidate preference to vary across horizons as future trajectories diverge and the relevance of available evidence changes over time. The framework combines native-time evidence modeling, candidate-time representations, relation-aware candidate–evidence reasoning, and signed evidential message passing to estimate horizon-specific reliability over a persistent bank of candidate futures. For oil-spill evolution, AMCR integrates simulator-derived candidates with SAR, optical, AIS, environmental, polygon, and source-related evidence, while maintaining a clear separation between upstream candidate construction and downstream evidence-guided reweighting. This design enables supportive, contradictory, and uncertain evidence to influence each candidate differently at each horizon. Experiments demonstrate improvements in spatial evolution estimation, candidate selection, localization, and calibration over simpler simulation, ensemble, and multimodal fusion baselines. The results further show that the most reliable candidate can change across prediction horizons, highlighting the importance of horizon-adaptive reasoning. Overall, AMCR provides a structured and interpretable mechanism for dynamically redistributing reliability across candidate futures according to the asynchronous multimodal evidence relevant to each target time.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, structural drafting, LaTeX preparation, and figure-layout prototyping. They were not used to create experimental measurements. All technical content, equations, references, and numerical results were independently verified by the authors.

REPRODUCIBILITY STATEMENT

The paper specifies the candidate interface, native-time evidence representation, factor-adaptive relations, evidential trust mechanism, candidate-time graph, optimization settings, evaluation metrics, and baseline definitions. The final release will include split manifests, cached candidate predictions, preprocessing code, random seeds, and scripts reproducing the reported tables and figures.

REFERENCES

Rami Al-Ruzouq, Mohamed Barakat A. Gibril, Abdallah Shanableh, Abubakir Kais, Osman Hamed, Saeed Al-Mansoori, and Mohamad Ali Khalil. Sensors, features, and machine learning for oil spill detection and monitoring: A review. *Remote Sensing*, 12(20):3338, 2020. doi: 10.3390/rs12203338.

C. J. Beegle-Krause. General NOAA oil modeling environment (GNOME): A new spill trajectory model. In *International Oil Spill Conference Proceedings*, pp. 865–871, 2001. doi: 10.7901/2169-3358-2001-2-865.

Filippo Maria Bianchi, Martine M. Espeseth, and Njal Borch. Large-scale detection and categorization of oil spills from SAR images with deep learning. *Remote Sensing*, 12(14):2260, 2020. doi: 10.3390/rs12142260.

Camilla Brekke and Anne H. Schistad Solberg. Oil spill detection by satellite remote sensing. *Remote Sensing of Environment*, 95(1):1–13, 2005. doi: 10.1016/j.rse.2004.11.015.

Fang Chen, Heiko Balzter, Feixiang Zhou, Peng Ren, and Huiyu Zhou. DGNet: Distribution guided efficient learning for oil spill image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. doi: 10.1109/TGRS.2023.3240579.

Knut-Frode Dagestad, Johannes Röhrs, Øyvind Breivik, and Bjørn Ådlandsvik. OpenDrift v1.0: A generic framework for trajectory modelling. *Geoscientific Model Development*, 11:1405–1420, 2018. doi: 10.5194/gmd-11-1405-2018.

Michela De Dominicis, Nadia Pinardi, George Zodiatis, and Robin Lardner. MEDSLIK-II, a lagrangian marine surface oil spill model for short-term forecasting-part 1: Theory. *Geoscientific Model Development*, 6:1851–1869, 2013. doi: 10.5194/gmd-6-1851-2013.

Philippe Delandmeter and Erik van Sebille. The Parcels v2.0 lagrangian framework: New field interpolation schemes. *Geoscientific Model Development*, 12:3571–3584, 2019. doi: 10.5194/gmd-12-3571-2019.

European Space Agency. Sentinel-1: Radar vision for Copernicus. ESA, Copernicus Programme; `https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-1`, 2026a. Accessed: Aug. 28, 2026.

European Space Agency. Sentinel-2: Colour vision for Copernicus. ESA, Copernicus Programme; `https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-2`, 2026b. Accessed: Aug. 28, 2026.

Jianchao Fan, Zitai Sui, and Xinzhe Wang. Multiphysical interpretable deep learning network for oil spill identification based on SAR images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024. doi: 10.1109/TGRS.2024.3357800.

Merv Fingas and Carl Brown. Review of oil spill remote sensing. *Marine Pollution Bulletin*, 83(1): 9–23, 2014. doi: 10.1016/j.marpolbul.2014.03.059.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1321–1330, 2017.

Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, Huimei He, Jian Wang, Jingdong Chen, Ming Yang, Yongjun Zhang, and Yansheng Li. SkySense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27672–27683, 2024.

Katerina Kikaki, Ioannis Kakogeorgiou, Ibrahim Hoteit, and Konstantinos Karantzalos. Detecting marine pollutants and sea surface features with deep learning in sentinel-2 imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 210:39–54, 2024. doi: 10.1016/j.isprsjprs.2024.02.017.

Marios Krestenitis, Georgios Orfanidis, Konstantinos Ioannidis, Konstantinos Avgerinakis, Stefanos Vrochidis, and Ioannis Kompatsiaris. Oil spill identification from satellite images using deep neural networks. *Remote Sensing*, 11(15):1762, 2019. doi: 10.3390/rs11151762.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6402–6413, 2017.

Duong Nguyen, Rodolphe Vadaine, Guillaume Hajduch, René Garello, and Ronan Fablet. GeoTrackNet-a maritime anomaly detector using probabilistic neural network representation of AIS tracks and a contrario detection. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):5655–5667, 2022. doi: 10.1109/TITS.2021.3055614.

Peng Ren, Qilin Jia, Qing Xu, Yongqing Li, Fan Bi, Jiangling Xu, and Song Gao. Oil spill drift prediction enhanced by correcting numerically forecasted sea surface dynamic fields with adversarial temporal convolutional networks. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–18, 2025. doi: 10.1109/TGRS.2025.3528631.

Stine Skrunes, Camilla Brekke, and Torbjorn Eltoft. Characterization of marine surface slicks by Radarsat-2 multipolarization features. *IEEE Transactions on Geoscience and Remote Sensing*, 52 (9):5302–5319, 2014. doi: 10.1109/TGRS.2013.2287916.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. EarthGPT: A universal multimodal large language model for multisensor image comprehension in remote sensing domain. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–20, 2024. doi: 10.1109/TGRS.2024.3409624.

A Appendix

This appendix provides supplementary details that complement the main AMCR formulation. We first describe the upstream construction of persistent oil-spill candidates from the observed spill state using PVD-AGN, which supplies the candidate bank before evidence-guided reweighting. We then present additional qualitative visualizations illustrating candidate evolution and multi-horizon spill predictions.

A.1 Upstream Candidate Construction Before AMCR Reweighting

The main paper treats candidate generation as external to AMCR. For the oil-spill instantiation, we use the candidate-construction stage of the Polygon Variational Drift-Attention Graph Network (PVD-AGN) as one concrete upstream realization of this interface. Its role is to transform an observed spill anchor and multiple physical configurations into a persistent bank of spatially comparable future hypotheses. The candidate bank is fixed before AMCR reweighting; no PVD-AGN reliability weight is passed to AMCR. AMCR therefore remains responsible for the horizon-dependent evidence-guided weights $W = [w_{k,h}]$.

From an observed spill to physical candidate futures. Let $\hat{M}_j$ denote the spill-like mask observed for scene j at anchor time T_0. The initial spill geometry is represented by

$$\mathcal{B}_j^0 = \mathrm{Poly}(\hat{M}_j). \tag{33}$$

Starting from this anchor, candidate k is generated under a distinct release, forcing, or simulator configuration,

$$\mathcal{Y}_j^{(k)} = \mathcal{D}_{\mathrm{spill}}\Big(\mathcal{B}_j^0, \mathbf{u}_j, \eta_j^{(k)}, \Delta t; \vartheta_j^{(k)}\Big), \qquad k = 1, \dots, K, \tag{34}$$

where $\mathbf{u}_j$ denotes environmental forcing, $\eta_j^{(k)}$ contains release/oil assumptions, and $\vartheta_j^{(k)}$ specifies the simulator and forcing perturbation. In our implementation, OpenDrift/OpenOil provides the physical trajectory path (Dagestad et al., 2018). Different configurations are retained as separate hypotheses rather than averaged during generation.

Raw particle trajectories are converted by a deterministic adapter into a common spatial and geometric representation,

$$\Gamma_j^{(k)} = \Psi(\mathcal{Y}_j^{(k)}) = \Big(P_{j,k}^{\mathrm{spill}}, P_{j,k}^{\mathrm{origin}}, \mathcal{B}_{j,k}^{+}, \mathcal{B}_{j,k}^{-}, \mathbf{d}_{j,k}^{\mathrm{oil}}, v_{j,k}^{\mathrm{oil}}, \Sigma_{j,k}^{\mathrm{evo}}\Big). \tag{35}$$

Here $P_{j,k}^{\mathrm{spill}}$ is the forward spill-probability field, $P_{j,k}^{\mathrm{origin}}$ is the backward release-likelihood field, $\mathcal{B}_{j,k}^{+/-}$ are forward/backward boundaries, $(\mathbf{d}_{j,k}^{\mathrm{oil}}, v_{j,k}^{\mathrm{oil}})$ summarize drift, and $\Sigma_{j,k}^{\mathrm{evo}}$ records the spatial evolution envelope. This standardization makes candidate differences reflect alternative physical hypotheses rather than incompatible simulator-output formats.

For the requested target times $\mathcal{T} = \{t_1, \dots, t_H\}$, the persistent AMCR candidate is obtained by retaining the same simulator identity across horizons,

$$Y_{k,h} = P_{j,k}^{\mathrm{spill}}(\cdot, t_h), \qquad C_k = \{Y_{k,h}\}_{h=1}^{H}, \qquad \mathcal{C} = \{C_k\}_{k=1}^{K}. \tag{36}$$

The geometric, drift, and evolution descriptors in $\Gamma_j^{(k)}$ are retained as candidate-side metadata for constructing the candidate-time representation. Thus the upstream path produces the fixed hypothesis bank $\mathcal{C}$, while AMCR separately receives asynchronous evidence $\mathcal{E}$ and estimates the horizon-specific reliability matrix W.

Polygon/evolution representation in PVD-AGN. PVD-AGN organizes each physical hypothesis through explicit polygon-variation and evolution tokens so that geometric deformation and forward transport are represented separately before being contextualized. For candidate k at prediction step j, a compact polygon-variation descriptor is

$$\nu_{j,k} = [\Delta V_{j,k}, \Delta A_{j,k}, \Delta c_{j,k}, \Delta\theta_{j,k}, \Delta\kappa_{j,k}, \sigma_{j,k}^{\mathrm{frag}}], \tag{37}$$

which summarizes displacement, area change, centroid drift, orientation change, circularity change, and fragmentation relative to the observed anchor. These quantities provide a compact description of how the candidate geometry departs from the spill polygon observed at T_0. In particular, centroid and orientation changes capture large-scale transport, while area, circularity, and fragmentation describe deformation of the spill footprint as the simulated trajectory evolves.

Candidate-specific polygon and evolution tokens are then formed as

$$\begin{aligned} V_{j,k}^{\mathrm{poly}} &= \phi_{\mathrm{poly}}(\nu_{j,k}), \\ T_{j,k}^{\mathrm{evo}} &= \phi_{\mathrm{field}}\Big(P_{j,k}^{\mathrm{spill}}, P_{j,k}^{\mathrm{origin}}, \Sigma_{j,k}^{\mathrm{evo}}\Big) + \phi_{\mathrm{dyn}}\big(\mathbf{d}_{j,k}^{\mathrm{oil}}, v_{j,k}^{\mathrm{oil}}\big) \\ &\quad + \phi_B(\mathcal{B}_{j,k}^{+}) + \phi_B(\mathcal{B}_{j,k}^{-}). \end{aligned} \tag{38}$$

Here, $V_{j,k}^{\mathrm{poly}}$ emphasizes the candidate's geometric variation, whereas $T_{j,k}^{\mathrm{evo}}$ summarizes its forward spill field, source/origin information, evolution uncertainty, drift displacement and velocity, and associated boundary information. The two representations are complementary: candidates with similar instantaneous spill masks may still differ in deformation history, drift direction, uncertainty, or source-related structure. Retaining these descriptors therefore preserves candidate diversity before AMCR evaluates which future is most consistent with the asynchronously available evidence.

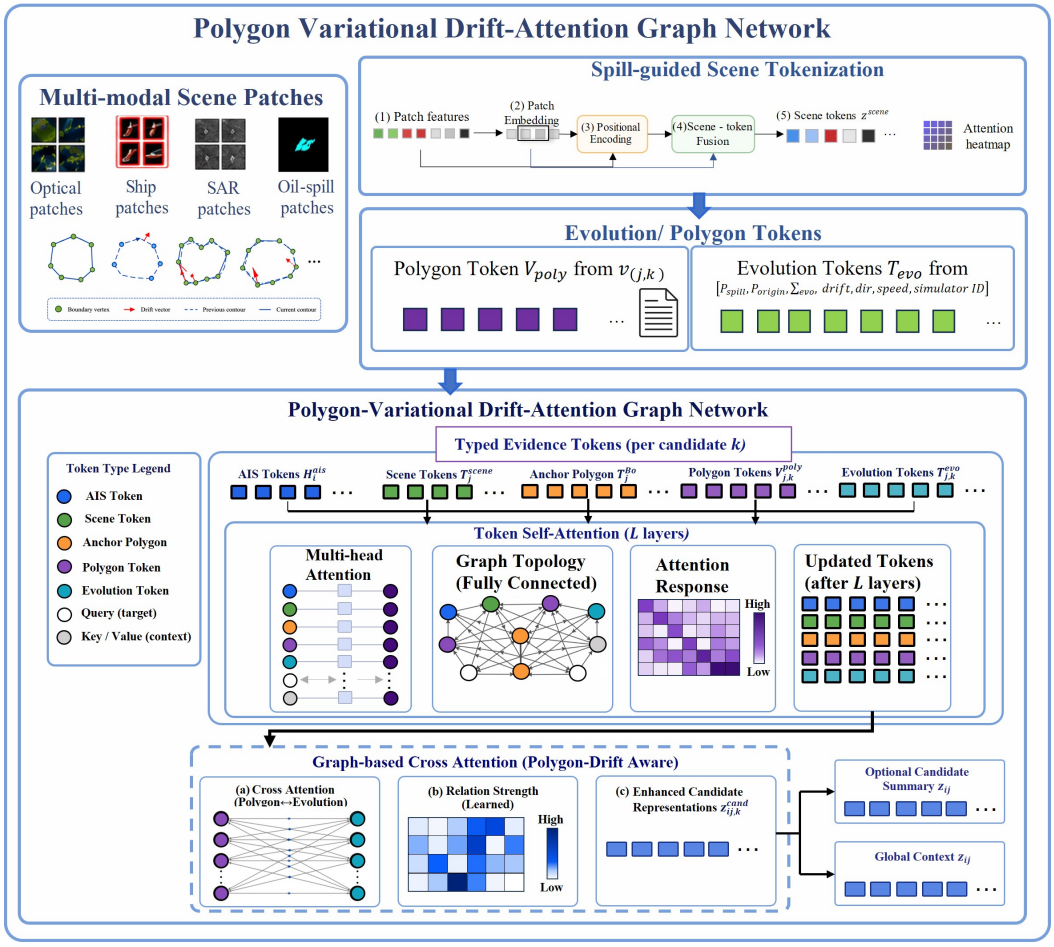

Figure 3: PVD-AGN candidate-construction and contextualization pipeline used to document the upstream origin of the oil-spill candidate bank. Observed spill geometry is converted into polygon/evolution representations and candidate-specific descriptors. For the AMCR integration, candidate states are fixed before AMCR and the PVD-AGN reliability weighting is not used.

As illustrated in Fig. 3, PVD-AGN further combines spill-guided scene tokenization with typed polygon and evolution tokens. Token self-attention provides contextual interaction among candidate-associated representations, while polygon–evolution cross-attention relates geometric variation to the corresponding forward evolution. The resulting representation is therefore not only a collection of independent simulated masks, but a structured description of how each persistent physical hypothesis evolves from the observed spill state.

The role of PVD-AGN in the present AMCR framework is deliberately restricted to this *upstream candidate-construction* stage. In particular, the handoff to AMCR occurs before PVD-AGN candidate-reliability weighting is applied. The persistent future states together with their polygon/evolution descriptors are retained as the candidate bank $\mathcal{C}$, while their reliability is recomputed by AMCR using native-time multimodal evidence, candidate–evidence relations, evidential polarity, and horizon-specific graph reasoning. Consequently, a candidate that is plausible from the upstream geometric and physical representation is not assumed to remain equally reliable at every prediction horizon. Its contribution can subsequently increase or decrease according to the evidence available for that target time.

This separation is important for the formulation in the main paper. PVD-AGN determines *which plausible futures are available for consideration*, whereas AMCR determines *how much each available future should contribute at each horizon*. The former therefore addresses candidate construction and representation, while the latter addresses asynchronous evidence-guided candidate-

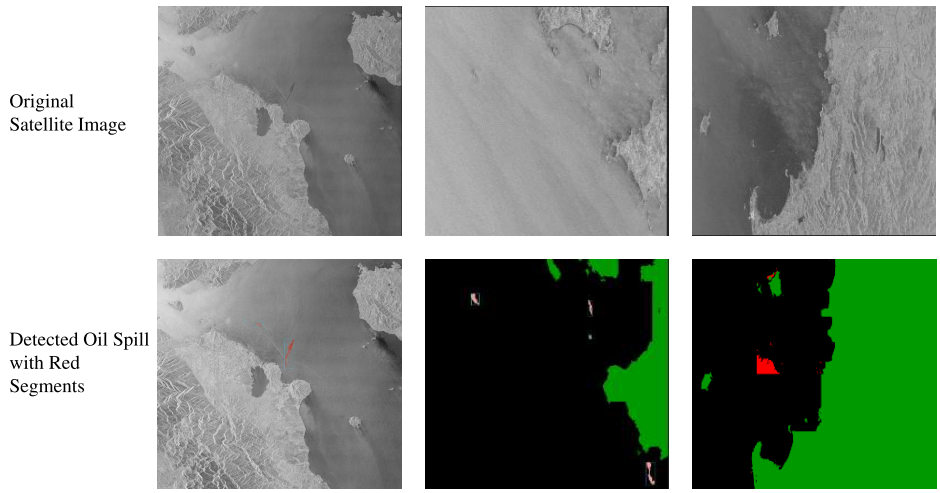

Figure 4: Representative Sentinel-1 SAR observations and detected oil-spill regions used at the observation/polygon-seed stage. The detected mask provides the anchor geometry for downstream simulator-candidate construction.

time reweighting. This also prevents the upstream candidate generator from implicitly absorbing the asynchronous evidence streams that AMCR is designed to reason over explicitly.

Candidate-generation configuration. All simulator outputs are mapped to a common 256×256 spatial grid before candidate representations are constructed. The cached simulator set varies three generation factors: particle count (500 or 2000), release-time offset (-1 or 0 h), and wind factor (0.9, 1.0, or 1.1). Their Cartesian combination produces 12 deterministic simulator configurations. These configurations provide controlled physical variation in the candidate bank while keeping the initial observed spill state fixed.

For the experiments reported in the main paper, the model-visible bank contains a fixed subset of six candidates in canonical generation order. The same candidate identity is preserved through all requested horizons, and forward states at 24, 48, and 72 h are retained. Thus,

$$C_k = \{Y_{k,24}, Y_{k,48}, Y_{k,72}\}, \qquad k = 1, \ldots, K, \tag{39}$$

with $K = 6$ for the model-visible bank. The persistent identity of C_k is essential for AMCR because reliability is estimated over candidate–time pairs (k, h) rather than over independently generated states at each horizon. A candidate can therefore receive high weight at one horizon and lower weight at another without changing its identity or replacing the underlying physical trajectory.

Candidate selection is completed before AMCR training and does not depend on ground-truth labels, learned AMCR scores, or validation performance. The resulting bank is cached and reused during training and evaluation. This keeps the upstream hypothesis set fixed so that improvements in the final prediction can be attributed to evidence-guided reweighting rather than to changes in candidate generation.

The fixed-bank design also separates two sources of forecasting quality. The first is *candidate coverage*: whether the upstream simulator and polygon-guided construction stage contain a sufficiently plausible future. The second is *candidate reasoning*: whether AMCR assigns appropriate horizon-dependent reliability to the available candidates. AMCR addresses the second problem and cannot recover a future that is absent from the candidate bank. This distinction motivates the oracle analysis in the main paper and provides a consistent interface between the upstream PVD-AGN construction stage and the downstream AMCR reweighting stage.

A.2 From Candidate Weighting to Asynchronous Candidate–Time Reweighting

Appendix A.1 describes how the fixed candidate bank is constructed before AMCR. We further clarify here why AMCR performs *candidate–time reweighting* rather than assigning one candidate weight and reusing it throughout the forecast window.

One-shot candidate weighting. A conventional candidate-weighting stage evaluates each physical hypothesis from the evidence available around the observed event and produces one normalized reliability value per candidate. PVD-AGN candidate-construction module combines scene agreement, physical plausibility, cross-candidate agreement, polygon consistency, uncertainty, and supportive or opposing evolution cues to produce a candidate score. Abstractly, this scoring process can be written as

$$\bar{s}_k = f_{\mathrm{w}}\left(C_k, \chi_k\right), \qquad \bar{w}_k = \frac{\exp(\bar{s}_k)}{\sum_{j=1}^{K} \exp(\bar{s}_j)}, \tag{40}$$

where χ_k denotes the candidate-specific visual, geometric, physical, and uncertainty cues. If these weights are reused for every prediction horizon, the fused prediction becomes

$$\widehat{Y}_h^{\mathrm{static}} = \sum_{k=1}^{K} \bar{w}_k Y_{k,h}. \tag{41}$$

This formulation allows candidate states $Y_{k,h}$ to evolve, but assumes that their relative reliability remains unchanged.

In AMCR, the handoff from the upstream candidate generator occurs *before* such reliability weighting. The PVD-AGN weights are therefore not used as priors or inputs to AMCR. The term *reweighting* instead refers to repeatedly redistributing reliability over the same persistent candidate identities as the target horizon and asynchronous evidence context change.

Horizon-conditioned reweighting. For each target t_h, AMCR first determines which evidence is causally available,

$$a_{i,h} = a_i \mathbb{I}[\tau_i^{\mathrm{obs}} \le t_h], \tag{42}$$

and then measures its target-dependent temporal relevance,

$$g_{i,h} = a_{i,h} \exp[-d_{m_i}(\tau_i^{\mathrm{occ}}, t_h)]. \tag{43}$$

The resulting evidence context is related independently to every candidate-time state $Y_{k,h}$ through relation-aware modeling and signed evidence propagation. This produces

$$s_{k,h} = F_{\mathrm{AMCR}}\left(C_k, Y_{k,h}, \mathcal{E}_h\right), \qquad w_{k,h} = \frac{\exp(s_{k,h})}{\sum_{j=1}^{K} \exp(s_{j,h})}, \tag{44}$$

where $\mathcal{E}_h$ denotes the evidence that is available and relevant to target t_h.

The distinction can therefore be summarized as

$$\underbrace{\bar{w}_k}_{\text{one weight per candidate}} \qquad \longrightarrow \qquad \underbrace{\{w_{k,1}, \ldots, w_{k,H}\}}_{\text{candidate–time reliability}}. \tag{45}$$

Candidate identity remains fixed; only its reliability is allowed to change.

Causal and temporal justification. The two-time evidence representation gives a direct justification for this reweighting procedure. If an evidence item has not yet become observable at target t_h, i.e., $\tau_i^{\mathrm{obs}} > t_h$, then

$$a_{i,h} = 0 \quad \Longrightarrow \quad g_{i,h} = 0 \quad \Longrightarrow \quad E_{i,h} = 0. \tag{46}$$

The masked graph attention therefore prevents that item from affecting $s_{k,h}$ or $w_{k,h}$. Once the item becomes available, its contribution is controlled by

$$g_{i,h} = \exp[-d_{m_i}(\tau_i^{\mathrm{occ}}, t_h)], \tag{47}$$

up to the availability factor. Hence the same observation may become available at one horizon and subsequently become less relevant as the target moves farther from its physical occurrence time.

Together with the bounded-influence result in Eq. 26, this yields the following procedure: the candidate bank is kept fixed, available evidence is selected using τ_i^{obs}, its relevance is evaluated using τ_i^{occ}, candidate–evidence relations are recomputed for the requested target, and the resulting candidate scores are renormalized across k. The procedure is repeated for every t_h.

Why reweighting is necessary. Suppose two horizons t_h and t_{h+1} share the same candidate identities but differ in either candidate state, evidence availability, or temporal relevance. Then the corresponding score vectors $s_{\cdot,h}$ and $s_{\cdot,h+1}$ need not coincide. Reusing a single $\bar{w}_k$ would suppress this change by assumption, whereas horizon-wise normalization permits

$$w_{k,h} \neq w_{k,h+1} \tag{48}$$

without changing the underlying physical candidate C_k. Reweighting therefore separates the question of *which futures are available* from the question of *how reliable each available future is at a particular target time*.

This distinction is also reflected by the AMCR-Static control in the main experiments. Both AMCR-Static and the full model use the same candidate universe and evidence sources; the difference is whether reliability is constrained to remain horizon-invariant or is estimated at the candidate–time level. The resulting performance gap therefore provides empirical support, within the evaluated setting, for allowing reliability to change across prediction horizons.

Optimization and implementation details. All candidate, evidence, relation, and variation features are projected to a shared dimension $d = 128$. Relation attention uses four heads with hidden width 128, candidate-time fusion uses gated projection and MLP refinement, and the reliability network employs a three-way support/contradiction/ignorance head. Signed evidence messages are aggregated before a two-layer score head of width 256. GELU, layer normalization, and dropout 0.1 are used throughout.

Temporal scales σ_m are Softplus-parameterized and initialized from modality-specific median inter-event intervals. We use $\lambda_T = 1$, $\lambda_0 = 0.1$, $\kappa = 1$, and $\zeta = 1$. Training combines fused-field supervision, per-horizon candidate-ranking supervision when available, the adaptive stability regularizer in Eq. 30, and calibration regularization. AdamW is used with learning rate 3×10^{-4}, weight decay 10^{-4}, gradient clipping at 1.0, and early stopping on the validation objective. All baselines use the same candidate cache, data split, training budget, and evaluation references.

A.3 Additional Qualitative Results and Visualizations

The following panels provide supplementary visual evidence for the observation-to-candidate interface and multi-horizon spill evolution. They complement the quantitative results in the main paper by illustrating how different initial observations and candidate evolutions behave over 24, 48, and 72 h.

A.3.1 Additional multi-horizon qualitative examples

Figures 5–8 provide additional qualitative examples of oil-spill evolution across 24, 48, and 72 h.

Figure 5 compares a single OpenDrift simulation with our candidate-based prediction for KOMPSAT-5 scene `K5_20240304212318_004039_57847_A_ST04_VV_SCS_B_L1A`, acquired on 4 March 2024. Across 24, 48, and 72 h, our prediction remains closer to the forward probability evolution, achieving Mask IoU values of 0.905/0.887/0.872 compared with 0.849/0.755/0.693 for the single OpenDrift simulation. The comparison illustrates the benefit of combining multiple candidate evolutions rather than relying on one fixed simulation trajectory.

Figure 6 provides a second multi-horizon example comparing forward probability evolution with a single OpenDrift realization and the corresponding candidate-based interpretation. The probability maps preserve multiple spatial modes as the spill expands and fragments, while the candidate representation consolidates the dominant evolution patterns without completely suppressing secondary regions. This illustrates why maintaining multiple plausible physical alternatives is useful before AMCR performs horizon-specific evidence-guided reweighting.

Figures 7 and 8 further show two MADOS optical examples (Kikaki et al., 2024) at 10 m ground-sampling distance. Each panel contains the optical observation at T_0, the detected spill mask, and the corresponding 24/48/72 h probability-evolution visualizations. Scene 133 exhibits a relatively elongated slick geometry, whereas Scene 150 contains a more spatially fragmented pattern. Despite these differences, the predicted fields remain spatially associated with the observed anchor while progressively changing their extent, shape, and spatial concentration over longer horizons.

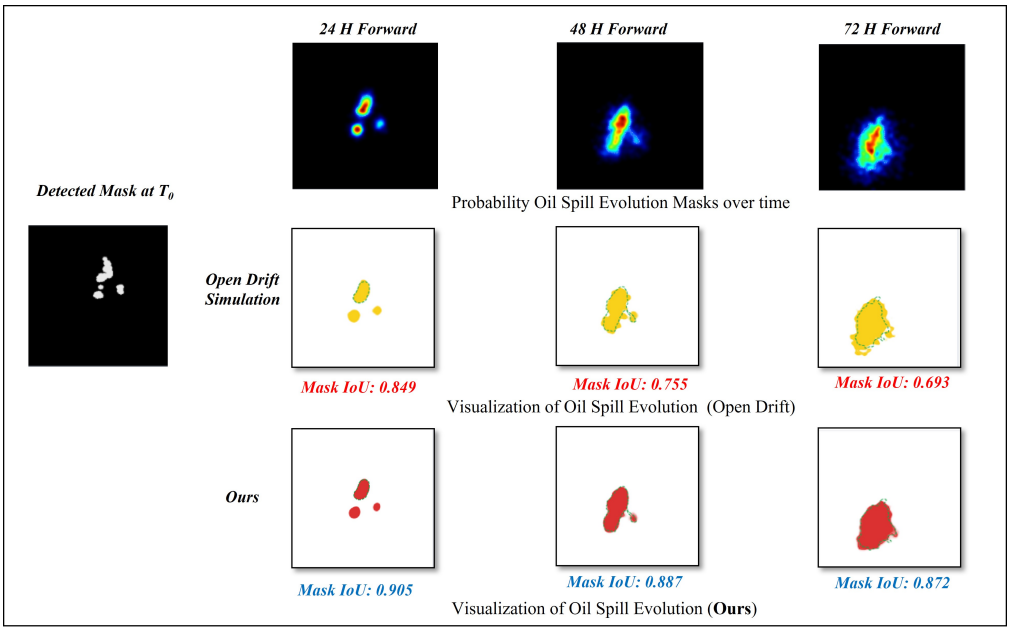

Figure 5: Forward oil-spill evolution for a KOMPSAT-5 scene acquired on 4 March 2024. Compared with a single OpenDrift simulation (Dagestad et al., 2018), our candidate-based prediction better follows the probability-evolution reference at 24, 48, and 72 h, achieving Mask IoU of 0.905/0.887/0.872 versus 0.849/0.755/0.693.

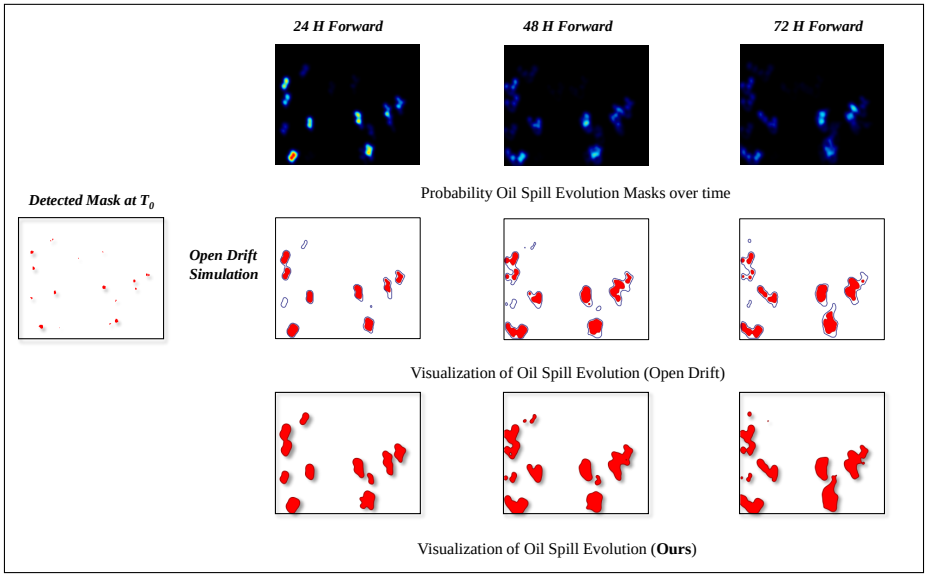

Figure 6: Additional oil-spill evolution example at 24, 48, and 72 h. The panel shows the detected anchor mask, forward probability evolution, a single OpenDrift simulation, and the candidate-based interpretation across the same horizons.

Together, Figures 5–8 illustrate the diversity of initial spill configurations and the resulting multi-horizon evolution patterns. They also emphasize the separation between the observed state at T_0, the set of plausible candidate futures, and the subsequent AMCR reweighting stage that determines how much each persistent candidate contributes at each prediction horizon.

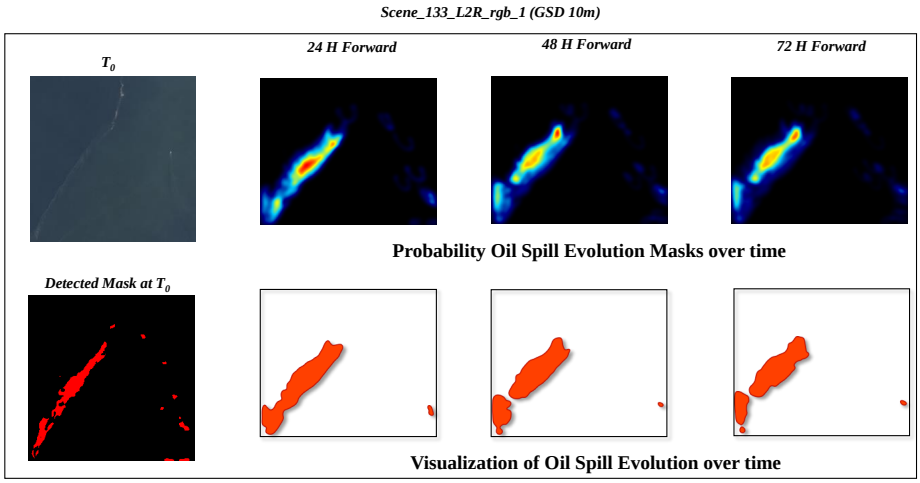

Figure 7: MADOS Scene 133 (10 m GSD): optical observation and detected mask at T_0, followed by 24/48/72 h probability-evolution visualizations.

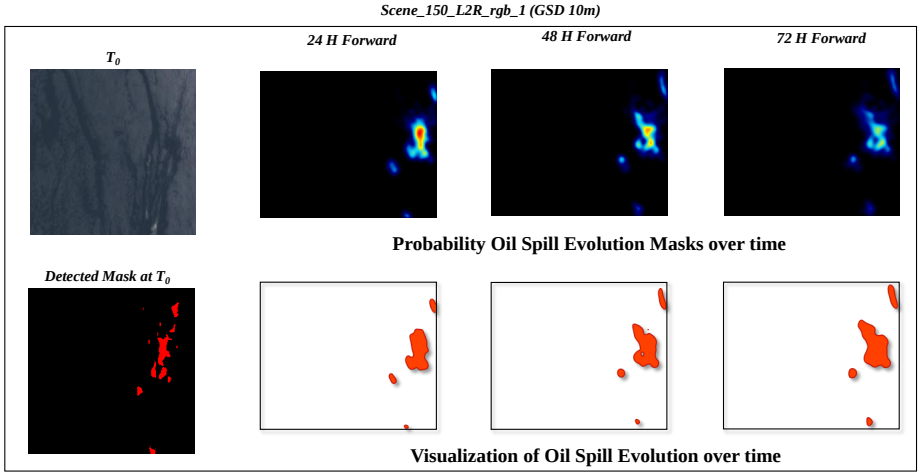

Figure 8: MADOS Scene 150 (10 m GSD): optical observation and detected mask at T_0, followed by 24/48/72 h probability-evolution visualizations.

The MADOS optical panels are included as qualitative observation-to-evolution examples only. They are not treated as temporally aligned forward ground-truth sequences and are therefore not used to compute the aggregate candidate-ranking, calibration, or fused-prediction results reported for AMCR.