SET PREDICTION FOR NEXT-DAY ACTIVE FIRE FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate next-day active fire forecasts can support early warning, disaster response, forest risk assessment, and downstream estimation of fire-related carbon emissions. Beyond these operational uses, accurate forecasting can provide a foundation for investigating which environmental factors are most informative about subsequent fire activity. Existing machine learning approaches to wildfire forecasting typically predict wildfire danger or fire probability on kilometre-scale daily grids, which is useful for regional warning but does not directly represent localised fire events. We propose Wildfire Ignition Set Predictor (WISP), a query-based model that reformulates next-day active fire forecasting as point-set prediction. From 48 hours of covariates including weather, vegetation products, static land, and fire history, WISP predicts a fixed-size ranked set of future active fire cluster centres on a $375\,\mathrm{m}$ grid across globally distributed regions. The model is trained end-to-end with Hungarian matching between predicted queries and observed fire-cluster centres, directly aligning supervision with the sparse, unordered structure of future fire activity. We construct a globally distributed, hourly, multi-source benchmark for this task. On a held-out test set spanning fire regions worldwide, WISP achieves the highest cluster recall among the evaluated pipelines, together with the highest localisation and coverage scores. The analysis based on fire regime identifies newly observed activity and false alarms following historical fire as key challenges. These results establish sparse set prediction as a viable formulation for high-resolution wildfire forecasting and provide a benchmark for future work in this domain.

1 INTRODUCTION

Wildfires displace communities, damage ecosystems, and impose growing economic and human costs, with the frequency of the most extreme events more than doubling globally between 2003 and 2023 (Cunningham et al., 2024). Fire-weather seasons have also lengthened across many regions (Jolly et al., 2015), increasing the need for anticipatory information. Operational systems such as the Global ECMWF Fire Forecast provide daily gridded fire-danger information to support early warning and proactive coordination of response resources (Di Giuseppe et al., 2020). Within this broader context, anticipating where active fire is likely to occur over the next day at fine spatial scale—rather than only identifying regions of elevated danger—is an operationally relevant prediction problem in Earth-system science. Once a model can anticipate future fire activity from heterogeneous environmental covariates, attribution and perturbation analyses can quantify which observed factors the model relies on, providing hypotheses about wildfire drivers (Kondylatos et al., 2026).

Predicting fire events at sub-kilometre scale is intrinsically difficult. Active fire occurrences are sparse, stochastic, and geographically diffuse, and depend on interactions among weather, fuel state, terrain, and human or natural ignition sources (Di Giuseppe et al., 2025). Statistical wildfire-occurrence models have therefore often represented fires as spatial or spatio-temporal point patterns (Koh et al., 2023; Opitz et al., 2020). At the nominal $375\,\mathrm{m}$ scale of VIIRS observations, a dense pixel-wise formulation is additionally exposed to severe foreground–background imbalance, a well-documented difficulty in dense detection and localization (Lin et al., 2020; Oksuz et al., 2021). These considerations motivate representing the next-day target directly as a sparse, unordered set of localised fire events. Satellite active fire products provide the observations needed to define such targets and are widely used for global fire monitoring and operational applications (Wooster et al., 2021). In particular, the VIIRS imaging-band active fire product provides observations at a nominal

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.

spatial resolution of $375\,\mathrm{m}$ (Schroeder et al., 2014). We use these observations to define the target reference grid and aggregate detections over the next 24 hours into active fire cluster centres.

We present Wildfire Ignition Set Predictor (WISP), which casts next-day active fire forecasting as sparse set prediction. Following DETR-type models (Carion et al., 2020; Song et al., 2021), WISP predicts a fixed-size set of query hypotheses, each comprising a fire probability and a normalized location; bipartite Hungarian assignment matches predictions to observed fire-cluster centres during training (Kuhn, 1955). WISP forecasts these centres from 48 hours of multi-source Earth observation and reanalysis covariates, including meteorology (Hersbach et al., 2020), satellite-derived vegetation (Fuster et al., 2020), static land descriptors (Jasiewicz & Stepinski, 2013; Nobre et al., 2011), and active fire history (see Section 4). Rather than predicting a scalar probability for every cell of a coarse forecast grid, WISP outputs a fixed budget of ranked fire-query hypotheses, each with a fire score and a normalized location.

The contributions of this work are three-fold. First, we formulate next-day active fire forecasting as point-set prediction and introduce WISP, which predicts a ranked set of localised future fire-cluster centres. Second, we design a multi-scale architecture that combines temporally resolved historical context, auxiliary land and fire representations, and gated future-weather conditioning with a reference-point query decoder. Third, we construct a globally distributed, hourly, multi-source benchmark and an evaluation protocol that distinguishes confidence-ranked detection, matched-cluster coverage and cluster recall. In the main baseline comparison, WISP improves cluster coverage and recovery, while WISP-OP retains strong performance without future weather inputs.

2 Related work

Machine learning for wildfire forecasting. Machine learning has been applied to several wildfire-related tasks, including danger forecasting (Apostolakis et al., 2022; McNorton et al., 2024), spread prediction (Huot et al., 2022; Shadrin et al., 2024; Rösch et al., 2024), burned-area probability and mapping (Son et al., 2024), and post-fire assessment. The line most relevant here, however, is short-term wildfire forecasting. Prior work in this setting typically predicts a cell-wise danger score or fire probability on a regular spatio-temporal grid, using either tree-based models on engineered covariates or neural networks trained on a wildfire datacube (Di Giuseppe et al., 2025; Kondylatos et al., 2023; 2022; 2026). This literature has evolved from regional case studies to benchmark datasets and, more recently, to global forecasting systems. Kondylatos et al. (2022) forecast next-day wildfire danger in the Eastern Mediterranean and use xAI to analyse the contribution of key drivers. Mesogeos (Kondylatos et al., 2023) later standardized this setting with a 17-year Mediterranean datacube at 1 km $\times$ 1 day resolution, together with benchmark tracks for short-term danger forecasting and burned-area estimation given an ignition point (similar to Shadrin et al. (2024)). WildfireSpreadTS (WSTS) provides multi-modal time series for next-day wildfire spread prediction on a $375\,\mathrm{m}$ grid (Gerard et al., 2023). Lahrichi et al. (2026) subsequently introduce WSTS+ and report UTAE(Res18) as their best-performing model on WSTS. These approaches forecast dense active fire maps, whereas WISP directly predicts a sparse set of future fire-cluster centres. More recently, Di Giuseppe et al. (2025) demonstrated global data-driven prediction of fire activity by combining weather, fuel, and ignition information. These studies establish the value of data-driven wildfire forecasting and treat the prediction primarily as dense gridded field estimation.

Set prediction for sparse localization. Set prediction is a natural framework when the target is a sparse, unordered set of entities. DETR (Carion et al., 2020) introduced set prediction with learned object queries and bipartite Hungarian matching (Kuhn, 1955) for object detection. Subsequent DETR-style work improved convergence, multi-scale feature use, query design, and matching stability (Zhu et al., 2021; Liu et al., 2022; 2023). This formulation has since been used beyond standard box detection. P2PNet (Song et al., 2021) directly predicts crowd points and matches them to head annotations with Hungarian assignment; Oriented DETR extends query-based set prediction to oriented objects in aerial imagery through a point-axis representation (Zhao et al., 2025b); and MapTR (Liao et al., 2023) represents vectorized map elements as structured point sets with hierarchical bipartite matching. These examples show that set prediction is well suited to sparse spatial outputs. WISP brings this sparse localization paradigm to next-day active fire forecasting.

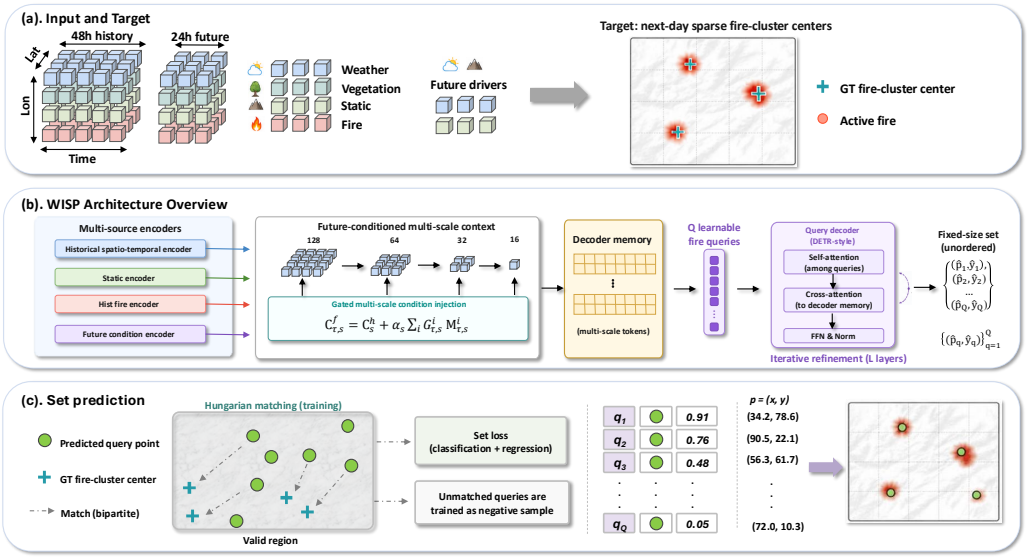


Fig. 1: Overview of WISP. Future-conditioned decoder memory supports reference-point query prediction of fire scores and cluster locations with Hungarian set supervision.

3 WISP-BENCHMARK

Table 1: Comparison with representative machine-learning wildfire benchmarks. Spatial resolution refers to the common prediction or target grid. ‡CONUS denotes the contiguous United States.

Dataset	Coverage	Period	Spatial	Temporal	Features*	Reported size†	Forecast target
Next Day Wildfire Spread (Huot et al., 2022)	CONUS	2012–2020	1 km	Daily	12	18,545 samples	Dense next-day spread
Mesogeos (Kondylatos et al., 2023)	Mediterranean	2006–2022	1 km	Daily	27	25,722 danger samples	Danger / burned area
WildfireSpreadTS (Gerard et al., 2023)	U.S.	2018–2021	375 m	Daily	23	607 fires 13,607 images	Dense next-day spread
SeasFire (Kondylatos et al., 2026)	Global	2001–2021	0.25°	8-day	59	966 global 8-day steps	Sub-seasonal fire patterns
TS-SatFire (Zhao et al., 2025a)	CONUS‡	2017–2021	375 m	Daily	27	179 fires 3,552 images	Detection / next-day progression
WISP-benchmark	**Globally distributed**	**2014–2024**	**375 m**	**Hourly**	**55**	**4,085 entites**	**Sparse next-day clusters**

We construct an 11-year active-fire benchmark spanning 2014–2024 by integrating Earth observation and reanalysis products on a common 375 m hourly grid. Each year, three $3^\circ \times 3^\circ \times 15$-day cubes are sampled from each of 14 GFED regions (Chen et al., 2023), yielding entities with 34 weather, 5 vegetation, 14 static-land, and 2 fire-observation variables. After quality filtering, the dataset contains 2,863 training, 406 validation, and 816 test entities; approximately 20% contain no future fire and 25% contain more than 20 active-fire pixels. Further details are provided in Appendix E. As summarised in Table 1, WISP-benchmark combines globally distributed sampling, 375 m spatial resolution, hourly temporal resolution, and 55 multi-source features over 11 years. Across existing benchmarks, global coverage is generally paired with coarser spatial or temporal resolution, whereas sub-kilometre datasets are predominantly regional and formulated as dense spread or detection tasks. In contrast to benchmarks centred on regional dense spread prediction or coarser global forecasting, WISP-benchmark directly targets sparse next-day active-fire clusters across diverse fire regimes.

4 WISP: WILDFIRE IGNITION SET PREDICTOR

4.1 PROBLEM FORMULATION

Forecast task. We formulate next-day wildfire active fire prediction as conditional fire cluster set prediction (see Fig. 1). Given historical Earth-system states $\mathbf{X}^h$ and future-known covariates $\mathbf{X}^f$, the goal is to predict a variable-length set of future fire-cluster centres over the next T_p hours. Rather than extrapolating past fire observations alone, the objective is to estimate the conditional distribution

of future fire-cluster locations given historical Earth-system states and future-known features:

$$P_\theta \left(\mathcal{G}_{T_p} \mid \mathbf{X}^h, \mathbf{X}^f \right), \tag{1}$$

where θ denotes the learnable parameters; $\mathcal{G}_{T_p}$ denotes the set of active fire clusters over the next T_p hours; $\mathbf{X}^h$ is the historical context up to time t, and $\mathbf{X}^f$ contains future-known covariates over the prediction horizon. The future fire field is not determined by historical fire observations alone, but by their interaction with meteorology, vegetation state, static land properties, and expected future weather.

Input Entity. The input to the model is a preprocessed multi-source entity (see (a) in Fig. 1):

$$\mathbf{X} \in \mathbb{R}^{F \times T \times H \times W}, \quad \mathbf{X} = \{\mathbf{W}, \mathbf{V}, \mathbf{S}, \mathbf{A}\}, \tag{2}$$

where $F = 55$ denotes the number of input channels, $T = 72$ is the temporal length, and $H = W = 128$ is the spatial context size. The entity contains four groups of variables: $\mathbf{W}$ contains 34 weather variables (e.g., temperature), $\mathbf{V}$ contains 5 vegetation-related features (e.g., leaf area index), $\mathbf{S}$ contains 14 static variables (e.g., elevation), and $\mathbf{A}$ contains 2 fire observations variables: active fire and fire radiative power (FRP). The full feature list is in Appendix A. Each entity is divided in to hist $T_h = 48$ input $\mathbf{X}^h$ and future $T_p = 24$ input $\mathbf{X}^f$:

$$\mathbf{X}^h = \{\mathbf{W}_{t-T_h+1:t}, \mathbf{V}_{t-T_h+1:t}, \mathbf{S}, \mathbf{A}_{t-T_h+1:t}\}, \quad \mathbf{X}^f = \left\{\mathbf{W}_{t+1:t+T_p}, \mathbf{S}\right\}, \tag{3}$$

where $\mathbf{W}$ is used in the historical and future windows; $\mathbf{V}$ is used only in the history; $\mathbf{S}$ is encoded once as a time-independent land prior. For encoding fire history, WISP derives a fire mask, a fire-recency feature, and two summary maps indicating historical fire occurrence and recency. These channels form the historical fire representation $\mathbf{A}^h_{\text{hist}}$ and summary fire history map $\mathbf{B}^h_{\text{sum}}$ (see Appendix B).

Target. The target is a variable length set of future fire cluster centres. We aggregate the future active fire observations $\mathbf{A}_{t+1:t+T_p}$ into a temporal union mask $\tilde{\mathbf{A}}$. Nearby fire pixels are then grouped into connected components using a 7×7 footprint, equivalent to a Chebyshev radius of 3 pixels. Since the VIIRS active fire product has a nominal spatial resolution of $375\,\text{m}$ (Schroeder et al., 2014), this radius corresponds to approximately $1\,\text{km}$. This grouping prevents adjacent detections from the same fire region from being treated as independent fire-cluster targets:

$$\mathcal{G}_K = \{\mathbf{g}_k\}_{k=1}^K, \qquad \mathbf{g}_k \in [0,1]^2, \tag{4}$$

where $\mathbf{g}_k$ is the FRP-weighted normalized centre location of active fire cluster and K is the number of connected components in $\tilde{\mathbf{A}}$. During training, when $K > Q$, only the Q highest-ranked target components are retained for Hungarian matching. The model outputs a fixed-size query set $\hat{\mathcal{Y}} = \{(\hat{p}_q, \hat{\mathbf{y}}_q)\}_{q=1}^Q$, where $Q = 10$ is the default setting, $\hat{p}_q$ is the fire probability, and $\hat{\mathbf{y}}_q \in [0,1]^2$ is the predicted cluster location. Unused queries are trained as no-fire queries (see (c) in Fig. 1).

4.2 Model Architecture

Spatio-Temporal Encoder. After adding the derived fire-history channels, the historical spatio-temporal encoder in Fig. 1(b) encodes $\mathbf{X}^h$ into a historical context $\mathbf{Z}^h$:

$$\mathbf{Z}^h = \left[\mathbf{W}^h, \mathbf{V}^h, \mathbf{A}^h_{\text{hist}}, \mathbf{S}^{seq}, \mathbf{B}^{h,seq}_{\text{sum}}\right] \in \mathbb{R}^{T_h \times C_{in} \times H \times W}, \quad C_{in} = 59 \tag{5}$$

where $\mathbf{S}^{\text{seq}}$ and $\mathbf{B}^{h,\text{seq}}_{\text{sum}}$ denote $\mathbf{S}$ and $\mathbf{B}^h_{\text{sum}}$ broadcast along the temporal dimension T_h, respectively. The historical tensor $\mathbf{Z}^h$ is processed by a spatio-temporal backbone encoder (Gao et al., 2022). A stride-4 patch embedding reduces each 128×128 frame to a 32×32 latent grid and projects the input channels to an embedding dimension $d_e = 128$:

$$\mathbf{E}^h = \text{PatchEmbed}(\mathbf{Z}^h) + \mathbf{P}^h_{1:T_h}, \qquad \mathbf{E}^h, \mathbf{P}^h_{1:T_h} \in \mathbb{R}^{T_h \times d_e \times H' \times W'}, \tag{6}$$

where $\mathbf{P}^h_{1:T_h}$ denotes a learnable spatio-temporal positional embedding added before attention and $\mathbf{E}^h$ is then processed by space-time self-attention blocks, using cuboids of size $4 \times 8 \times 8$ and global vectors for cross-cuboid communication (Gao et al., 2022). Instead of collapsing the historical window into a single feature map, the backbone preserves a compact T_h-step historical sequence at 32×32 resolution. After the attention blocks, an output projection produces a 64-channel temporal representation, which is compressed by a history context projector. The projector first reduces each

temporal feature from 64 to 16 channels and then folds the temporal dimension into the channel dimension:

$$[T_h, 64, 32, 32] \;\rightarrow\; [T_h, 16, 32, 32] \;\rightarrow\; [T_h \cdot 16, 32, 32]. \tag{7}$$

The folded representation is further converted into a four-level historical context pyramid $\mathcal{C}^h = \{\mathbf{C}_s^h\}_{s\in\{128,64,32,16\}}$, where the s indicates the spatial resolution of each context level. These multi-scale historical features are used by the subsequent condition-injection modules.

Land and Fire Feature Encoding. In parallel to the historical spatio-temporal encoder, WISP encodes three auxiliary sources into scale-aligned spatial pyramids: the static land prior $\mathbf{S}$, the historical fire sequence $\mathbf{A}_{\text{hist}}^h$, and the historical fire-summary map $\mathbf{B}_{\text{sum}}^h$:

$$\Phi_{\text{land}}(\mathbf{S}) = \{\mathbf{S}_s\},\ \Phi_{\text{af}}\left(\mathbf{A}_{\text{hist}}^h\right) = \left\{\mathbf{A}_s^h\right\},\ \Phi_{\text{fire}}\left(\mathbf{B}_{\text{sum}}^h\right) = \left\{\mathbf{B}_s^h\right\},\ s \in \{128, 64, 32, 16\}. \tag{8}$$

The three encoders do not share parameters, and their outputs are aligned with the historical context pyramid $\mathcal{C}^h$. These pyramids provide scale-wise land and fire priors for the subsequent gated condition-injection module, as shown in Fig. 1(b). Architectural details are provided in Appendix C.

Future Condition Encoding. WISP then encodes the future drivers over the prediction horizon, as shown in Fig. 1(b). $\tau \in \{1, \ldots, T_p\}$, future weather is encoded into a four-level condition pyramid $\mathcal{W}_\tau^f = \{\mathbf{W}_{\tau,s}^f\}_{s\in\{128,64,32,16\}}$. To provide each future step with horizon-level context, WISP also encodes the T_p-step future features into one future-sequence context. The step-wise and future-sequence context features are added at each scale:

$$\bar{\mathbf{W}}_{\tau,s}^f = \mathbf{W}_{\tau,s}^f + \mathbf{W}_{\text{seq},s}^f, \quad s \in \{128, 64, 32, 16\}. \tag{9}$$

Thus, each prediction step receives both its local future feature and the global future trajectory.

Gated Multi-Scale Condition Injection. WISP constructs a conditioned context pyramid for each future step (Fig. 1(b)). For each scale s and forecast step τ, the model starts from the historical context feature $\mathbf{C}_s^h$ and injects four condition features: $\mathcal{M} = \{\mathcal{W}_\tau^f, \Phi_{\text{land}}(\mathbf{S}), \Phi_{\text{af}}\left(\mathbf{A}_{\text{hist}}^h\right), \Phi_{\text{fire}}\left(\mathbf{B}_{\text{sum}}^h\right)\}$.

Rather than concatenating all sources directly, WISP uses gated residual injection (Srivastava et al., 2015; Perez et al., 2018). At each scale, every source is first projected into the same channel space as the historical context. A source-specific gate then controls how much that source modifies the historical feature:

$$\mathbf{C}_{\tau,s}^f = \mathbf{C}_s^h + \alpha_s \sum_i \mathbf{G}_{\tau,s}^i \odot \mathcal{M}_{\tau,s}^i. \tag{10}$$

Here, $\mathcal{M}_{\tau,s}^i$ denotes the learned residual feature derived from i-th conditioning source in $\mathcal{M}$, $\mathbf{G}_{\tau,s}^i \in [0, 1]$ is its context-dependent gate, and α_s is a scale-dependent injection coefficient. The residual feature is computed from the projected source representation, whereas the gate is conditioned jointly on the historical context and the corresponding source feature. In WISP, the high-resolution features are updated more conservatively, while lower-resolution semantic features receive stronger conditioning (Tan et al., 2020): $\alpha_{128} = 0.25, \alpha_{64} = 0.50, \alpha_{32} = 1.00, \alpha_{16} = 1.00$. This yields a four-level future-conditioned pyramid $\mathcal{C}_\tau^f = \{\mathbf{C}_{\tau,s}^f\}_{s\in\{128,64,32,16\}}$ for each forecast step. This design preserves the multi-scale structure of the historical representation while allowing future forcing and land/fire priors to modulate it in a source-aware and scale-aware manner. The conditioned pyramid is then converted into a fixed-resolution decoder memory:

$$\mathbf{M}_\tau = \text{Flatten}\left(\Gamma_{\text{mem}}(\mathcal{C}_\tau^f)\right) \in \mathbb{R}^{1024\times 128}, \mathbf{M} = \left[\mathbf{M}_1, \ldots, \mathbf{M}_{T_p}\right]. \tag{11}$$

Here, Γ_{mem} projects the four pyramid levels to a common 32×32 memory grid with 128 channels, and $\text{Flatten}(\cdot)$ converts it into 1024 spatial tokens. The stacked memory $\mathbf{M}$ is used as the future-conditioned spatio-temporal context for the query decoder.

Query Decoding. Given the future-conditioned memory $\mathbf{M}$, WISP predicts a fixed-size set of future fire-cluster hypotheses using a query decoder (Fig. 1). The memory is first flattened over the prediction horizon and spatial tokens, from $\mathbf{M} \in \mathbb{R}^{T_p\times 1024\times 128}$ to $\mathbf{M}^\flat \in \mathbb{R}^{(T_p\cdot 1024)\times 128}$. A positional embedding is added to each memory token using its forecast step, row index, and column index. This allows the decoder to distinguish not only where a memory token is located, but also which future hour it represents. The decoder uses Q learnable queries. Each query has a content embedding $\mathbf{q}_q^0 \in \mathbb{R}^{128}$ and a learnable reference point $\mathbf{r}_q^0 \in [0, 1]^2$, for $q = 1, \ldots, Q$. The reference point $\mathbf{r}_q^0$ is defined in the normalized valid prediction region (see Section 5). WISP maps the reference

point back to the full context coordinates when constructing query positional embeddings. Following DETR-style query decoding and bipartite matching design (Carion et al., 2020; Liu et al., 2022), each decoder layer performs self-attention among queries, cross-attention to the memory tokens, and a feed-forward update. The query reference point is then refined by an additive offset in inverse-sigmoid space:

$$\mathbf{r}_q^\ell = \sigma\left(\sigma^{-1}(\mathbf{r}_q^{\ell-1}) + \Delta_q^\ell\right), \quad \ell = 1, \ldots, L. \tag{12}$$

In the current implementation, $L = 2$. This iterative update gives each query an explicit spatial hypothesis and allows the decoder to progressively move it toward a future fire-cluster centre. The final query states are mapped to fire classification logits and normalized point coordinates:

$$\hat{\mathcal{Y}} = \{(\hat{p}_q, \hat{\mathbf{y}}_q)\}_{q=1}^{Q}, \quad \hat{p}_q = \mathrm{softmax}(\mathbf{z}_q), \quad \hat{\mathbf{y}}_q = \mathbf{r}_q^L. \tag{13}$$

Here, $\hat{p}_q$ is the predicted fire probability and $\hat{\mathbf{y}}_q \in [0,1]^2$ is the predicted fire-cluster location in the valid region (see Equation 4).

Hungarian Set Objective. WISP is trained as a fixed-size set predictor. For each entity, the decoder outputs Q fire-cluster predictions, while the target is a variable-size set $\mathcal{G}_K = \{\mathbf{g}_k\}_{k=1}^{K}$. Following DETR-style set prediction (Carion et al., 2020) and the Hungarian matching algorithm (Kuhn, 1955), we first compute a one-to-one assignment between predicted queries and ground-truth fire clusters (see Fig. 1(c)). For query q and target cluster k, the matching cost is defined as

$$\mathcal{D}_{qk} = \lambda_{\mathrm{m,cls}}\left(-\hat{p}_q\right) + \lambda_{\mathrm{m,loc}} \left\|\hat{\mathbf{y}}_q - \mathbf{g}_k\right\|_1, \tag{14}$$

where $\hat{p}_q$ is the predicted fire probability and $\hat{\mathbf{y}}_q \in [0,1]^2$ is the predicted cluster location. The first term encourages a ground-truth cluster to be assigned to a confident fire query, whereas the second term favours spatially close predictions. The assignment is then obtained as $\pi^\star = \arg\min_\pi \sum_{(q,k)\in\pi} \mathcal{D}_{qk}$, where π ranges over one-to-one matchings between predicted queries and target clusters. This matching step determines which queries are treated as positive fire predictions and queries not selected by $\pi^\star$ are assigned to the no-fire class. An illustrative query-to-cluster cost matrix is provided in Appendix D. After matching, WISP applies classification supervision to all queries and localization supervision only to matched queries. Let $c_q^\star = 1$ if query q is matched to a target cluster and $c_q^\star = 0$ otherwise (see Hungarian Matching algorithm in Appendix D). The classification loss is

$$\mathcal{L}_{\mathrm{cls}} = -\frac{1}{Q}\sum_{q=1}^{Q}\left[c_q^\star \log \hat{p}_q + w_{\mathrm{eos}}(1 - c_q^\star)\log(1-\hat{p}_q)\right], \qquad w_{\mathrm{eos}} = 0.1 \tag{15}$$

where unmatched queries are assigned to the no-object class and down-weighted by $w_{\mathrm{eos}} = 0.1$. The classification loss is normalized by the fixed query count (BQ for a batch of B entities), rather than by the sum of class weights. This preserves the down-weighting even for entities without target clusters. The localization loss is computed only for matched query-target pairs:

$$\mathcal{L}_{\mathrm{loc}} = \frac{1}{2|\pi^\star|}\sum_{(q,k)\in\pi^\star} \left\|\hat{\mathbf{y}}_q - \mathbf{g}_k\right\|_1, \tag{16}$$

where the factor 2 normalizes over the two coordinate dimensions; and with $\lambda_{\mathrm{loc}} = 5.0$, the loss function is

$$\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{loc}}\mathcal{L}_{\mathrm{loc}}. \tag{17}$$

5 Experiments

Weather-input settings. We evaluate two input settings. WISP conditions on meteorological fields over the future 24-hour window, obtained from ERA5 reanalysis. WISP-OP excludes all weather inputs from this future window. The comparison assesses the value of future-weather information; evaluation with weather forecasts available at issue time remains a separate step towards deployment.

Baselines. We compare WISP with representative approaches spanning four modelling paradigms. To make their outputs comparable for active fire cluster localization, dense forecasts are converted into scored points using frozen extraction adapters, whereas point-set models directly predict localised hypotheses. The four groups comprise: (i) Fire Persistence, which forecasts next-day activity from

Table 2: Overall point-set evaluation on WISP benchmark test dataset (%). Red and blue denote the best and second-best results. ‡ denotes model-specific score-field-to-point readout.

Model	mAP	MassCov@14	Hit@14	ClusRec@14	Top10Rec@14
Fire Persistence	15.54	29.87	36.31	30.67	25.67
Earthformer‡	11.24	43.91	46.11	38.77	48.61
Galileo‡	11.32	39.53	41.23	35.23	42.71
Chronos-2‡	17.51	46.61	48.43	40.96	53.06
ClimaX‡	15.86	46.59	47.35	41.89	48.92
WSTS+‡	13.17	27.97	37.26	23.13	22.02
Marked ST-Hawkes	**26.13**	24.36	31.46	25.92	**59.88**
WISP-OP	22.41	**50.02**	**51.33**	**45.12**	58.89
WISP	**23.31**	**54.28**	**54.52**	**49.73**	**64.47**

the preceding day's observed fires; (ii) a statistical marked spatio-temporal Hawkes process (Xu et al., 2023); (iii) task-specific wildfire spread forecasting, represented by WSTS+ UTAE–Res18, adapted from the best-performing model on WSTS reported by Lahrichi et al. (2026); and (iv) general spatio-temporal and pretrained-model adaptations, including Earthformer–SEVIR (Gao et al., 2022), Galileo (Tseng et al., 2025), Chronos-2 (Ansari et al., 2025), and ClimaX with a historical-fire CNN (Nguyen et al., 2023). All pipelines are evaluated on the same 816 test dataset, without assuming identical inputs or training budgets.

Evaluation metrics. We report five complementary point-set metrics, with formal definitions provided in Appendix F. Event **mAP** averages AP at spatial tolerances of 7, 14, and 21 pixels, where 14 pixels correspond to 5.25 km. **MassCov@14** measures FRP-weighted matched-cluster coverage, while **Hit@14** measures the fraction of matched pairs within 14 pixels. **ClusRec@14** measures cluster recovery at the operating threshold, whereas **Top10Rec@14** measures threshold-free recovery of the strongest target clusters using the ten highest-scoring predictions. All metrics are reported as percentages, with higher values indicating better performance.

5.1 Forecasting Performance

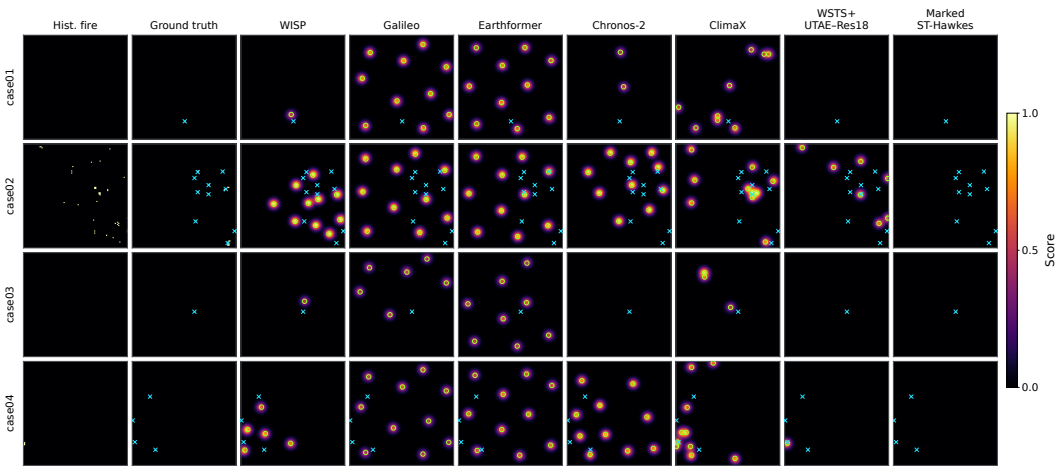


Fig. 2: Qualitative comparison on four test entities: cases 01–04. Columns show historical fire observations, ground-truth cluster centres over the next 24 hours, and model predictions. Cyan crosses mark ground-truth centres and circles mark predicted centres. The shared colour scale indicates prediction scores. Coloured blobs are renderings of point scores, not predicted fire perimeters or fire radiative power.

Main comparison. Table 2 shows that WISP achieves the highest MassCov@14, Hit@14, ClusRec@14 and Top10Rec@14 in the main comparison, at 54.28%, 54.52%, 49.73% and 64.47%, respectively. Its event mAP is 23.31%, second to Marked ST-Hawkes at 26.13%. Hawkes also

Table 3: WISP module and training-recipe comparisons on the test dataset (%). Red and blue denote the best and second-best results.

Model	mAP	MassCov@14	Hit@14	ClusRec@14	Top10Rec@14
A1: gates fixed to one	21.66	42.04	45.64	40.86	55.41
A2: no auxiliary paths	17.48	33.45	37.61	33.51	50.04
A3: temporal averaging	21.30	43.15	46.99	39.83	58.23
A4: R(2+1)D backbone	**23.62**	46.25	48.61	45.08	**60.16**
A5: no query SA	20.87	36.28	41.15	35.67	56.41
DETR–head	21.96	**48.89**	**51.47**	**47.08**	59.03
WISP	**23.31**	**54.28**	**54.52**	**49.73**	**64.47**

achieves a Top10Rec@14 of 59.88%, but lower coverage, hit rate and cluster recall. Thus, confidence-ranked detection and recovery of the issued prediction set yield different model rankings.

Compared with WSTS+ UTAE–Res18, WISP increases mAP from 13.17% to 23.31%, ClusRec@14 from 23.13% to 49.73%, and Top10Rec@14 from 22.02% to 64.47%. WISP also outperforms the evaluated Earthformer, Galileo, Chronos-2 and ClimaX adaptations on all five metrics. These results support WISP for the evaluated cluster-forecasting task; because inputs, training recipes and output readouts differ, they do not isolate output representation as the sole source of the performance gains.

Forecasting without future weather. WISP-OP retains strong performance without any future-weather inputs, achieving 22.41% mAP, 50.02% MassCov@14, 51.33% Hit@14, 45.12% ClusRec@14 and 58.89% Top10Rec@14. It outperforms the other neural forecasting adaptations in Table 2 on all five metrics. Weather-conditioned WISP improves these scores by 0.90, 4.26, 3.19, 4.61 and 5.58 percentage points, respectively. The larger gains in coverage and cluster recovery indicate that future-weather information provides additional predictive value, while competitive performance remains possible without it.

Qualitative results. Fig. 2 compares four selected test entities with different spatial configurations of future fire activity. In cases 01 and 03, no historical fire is visible in the displayed panels, and WISP places a single displayed prediction close to the future ground-truth centre. In contrast, the Galileo and Earthformer adaptations retain multiple spatially dispersed candidates, whereas WSTS+ UTAE–Res18 displays no candidate in these two cases. In case 02, WISP distributes its predictions across the region containing multiple target clusters, although some targets remain uncovered and some predictions are displaced. Case 04 illustrates remaining localisation errors: WISP concentrates its predictions in the lower-left region but retains off-target candidates. Chronos-2 and ClimaX show mixed behaviour, with predictions near some targets alongside additional or missing candidates. WSTS+ UTAE–Res18 produces multiple candidates in case 02 and a boundary-adjacent prediction in case 04. No predicted markers are visible for Marked ST-Hawkes in these examples. These selected cases illustrate candidate placement and activation under the display protocol, rather than dataset-level performance rankings. The coloured blobs visualise point scores, not predicted fire extent. Additional good, median, and failure cases are provided in Appendix I. For illustrating a clear qualitative convergence pattern, with initially dispersed activations progressively concentrating around the target region (Appendix H). The query location evolution shows a clear qualitative convergence pattern, with dispersed queries progressively concentrating around the target region (Appendix H).

5.2 Ablation Study

Architectural components. We examine five architectural changes relative to WISP. A1 fixes the source-specific conditioning gates to one, retaining the residual updates and scale-dependent injection coefficients. A2 removes the dedicated land, historical-fire and fire-summary pathways, while retaining these covariates in the historical backbone input. A3 replaces temporal-to-channel folding in the history context projector with temporal mean pooling. A4 replaces the historical Earthformer backbone with an R(2+1)D convolutional backbone (Tran et al., 2018). A5 removes self-attention among decoder queries while retaining memory cross-attention, feed-forward updates and reference-point refinement. Table 3 evaluates the contribution of WISP's architectural components. Fixing the conditioning gates to one reduces MassCov@14 from 54.28% to 42.04%. Removing the auxiliary paths produces the largest coverage and recall losses among A1–A5, with MassCov@14 and ClusRec@14

falling to 33.45% and 33.51%, respectively. Temporal averaging reduces Top10Rec@14 from 64.47% to 58.23%, while removing query self-attention reduces ClusRec@14 from 49.73% to 35.67%. These comparisons support the contribution of adaptive conditioning, auxiliary representations, temporal information retention and query interactions to cluster recovery. The R(2+1)D backbone achieves a slightly higher mAP than WISP (23.62% versus 23.31%), but lower scores on the other four metrics. WISP also outperforms the DETR-based comparator on all five metrics.

Query budget and matching cost. We further vary the query budget and Hungarian matching weights (Appendix G). Location-only matching performs best among the tested $Q = 10$ configurations. With matching weights fixed at $(1, 2)$, increasing Q from 10 to 20 improves mAP from 23.10% to 26.02% and ClusRec@14 from 43.75% to 58.66%, but reduces Top10Rec@14 from 61.50% to 59.28%. Thus, a larger query budget improves overall cluster recovery without necessarily improving recovery within the ten highest-ranked predictions. It also increases decoder computation and the storage required for query activations. Query budget and matching design therefore offer configurable trade-offs between recovery, output-budget constraints and computational requirements, to be selected and validated for the intended application.

5.3 Regime-wise Analysis

Table 4: Regime-wise performance on the test set. Fire-present metrics and Mean prob. are in %; Avg. pred. is a count. Red and blue denote the best and second-best results. ‡ denotes model-specific score-field-to-point readout.

	New ($n = 138$)			Continued ($n = 386$)			Extinguished ($n = 59$)		Quiescent ($n = 233$)	
Model	mAP	MassCov@14	Hit@14	mAP	MassCov@14	Hit@14	Avg. pred. ↓	Mean prob. ↓	Avg. pred. ↓	Mean prob. ↓
WSTS+‡	0.35	0.72	0.72	15.39	37.71	50.32	**0.98**	**0.42**	**0.03**	**0.01**
Chronos-2	8.28	**20.19**	**21.24**	18.79	56.06	58.15	9.73	4.64	2.03	2.52
WISP-OP	**13.54**	19.68	20.55	**23.94**	**60.87**	**62.34**	8.22	4.28	**0.78**	1.96
WISP	**12.59**	**24.96**	**25.49**	**24.75**	**64.76**	**64.90**	**7.54**	**3.94**	0.75	**1.57**

We further analyse WISP by separating the test set according to past-fire presence ($\mathrm{PF} = 1/0$) and future-fire presence ($\mathrm{FF} = 1/0$). This defines four observation-based regimes: **New** ($\mathrm{PF} = 0, \mathrm{FF} = 1$), **Continued** ($\mathrm{PF} = 1, \mathrm{FF} = 1$), **Extinguished** ($\mathrm{PF} = 1, \mathrm{FF} = 0$), and **Quiescent** ($\mathrm{PF} = 0, \mathrm{FF} = 0$). The four subsets contain 138, 386, 59, and 233 entities, respectively. For the fire-present regimes, we report mAP, MassCov@14, and Hit@14. For the two regimes without future observed fire, no target cluster is available for event localisation; we therefore report the average number of issued predictions per entity (Avg. pred.) and the mean value of the Gaussian-rendered prediction map (Mean prob.). Mean prob. is reported in percent and corresponds to the mean rendered score ($\times 100$) using the common $Q = 10$, $\sigma = 3$ rendering; it should not be interpreted as a calibrated probability of fire.

New versus continued activity. WISP performs substantially better when active fire is already observed in the historical window. From New to Continued cases, mAP increases from 12.59% to 24.75%, MassCov@14 from 24.96% to 64.76%, and Hit@14 from 25.49% to 64.90%. This indicates that newly observed activity without a historical fire cue remains substantially more difficult. Because the two regimes may also differ in event size, location, and environmental conditions. Future-weather conditioning improves spatial recovery in both fire-present regimes. This indicates that future-weather information provides additional spatial predictive value, while its effect on confidence-ranked detection is regime dependent.

6 Conclusion

We introduced WISP, a query-based model for next-day active fire cluster forecasting on the VIIRS 375 m reference grid. Across the evaluated forecasting baselines, WISP achieves the strongest matched-cluster coverage and top-ten recovery. Component comparisons support gated conditioning, auxiliary feature pathways, temporal information retention and query interactions, while also revealing trade-offs with confidence-ranked detection. WISP-OP retains strong performance without future weather inputs, and future-weather conditioning provides additional gains. Future work should improve score calibration and query suppression, evaluate issue-time weather forecasts, and extend the framework towards coupled fire occurrence and spread modelling.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing and grammar correction, to support the execution of baseline experiments, and to organize and format the final performance comparison tables. All AI-assisted code and experimental procedures were reviewed and verified by the authors, and all reported numerical results and table entries were checked against the original experimental outputs. Generative AI was not used to generate research data, formulate theoretical claims or proofs, or independently determine the scientific conclusions of this work. The authors reviewed all AI-assisted work and take full responsibility for the final content of the paper.

ETHICS STATEMENT

This study uses Earth observation and reanalysis data for next-day active fire forecasting and does not involve individual-level sensitive data or direct interaction with human participants. Potential downstream risks include missed fires and false alarms. Operational use would require further validation and should not rely on these forecasts as the sole basis for emergency decisions.

REPRODUCIBILITY STATEMENT

The input protocol, target construction, architecture and training objective are defined in Section 4. Data sources and preprocessing are described in Appendix A and Appendix E, and the matching algorithm is given in Appendix D. Section 3 and Appendix F describe the evaluation criteria and operating thresholds, distinguishing model-specific readouts from the common event evaluator. Section 5.3 defines the observation-based regime partition. We plan to release the implementation and derived benchmark data upon acceptance.

REFERENCES

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025. URL https://arxiv.org/abs/2510.15821.

Alexis Apostolakis, Stella Girtsou, Giorgos Giannopoulos, Nikolaos S. Bartsotas, and Charalampos Kontoes. Estimating next day's forest fire risk via a complete machine learning methodology. *Remote Sensing*, 14(5), 2022. ISSN 2072-4292. doi: 10.3390/rs14051222. URL https://www.mdpi.com/2072-4292/14/5/1222.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, pp. 213–229, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-58451-1. doi: 10.1007/978-3-030-58452-8_13. URL https://doi.org/10.1007/978-3-030-58452-8_13.

Y. Chen, J. Hall, D. van Wees, N. Andela, S. Hantson, L. Giglio, G. R. van der Werf, D. C. Morton, and J. T. Randerson. Multi-decadal trends and variability in burned area from the fifth version of the global fire emissions database (gfed5). *Earth System Science Data*, 15(11):5227–5259, 2023. doi: 10.5194/essd-15-5227-2023. URL https://essd.copernicus.org/articles/15/5227/2023/.

Calum X Cunningham, Grant J Williamson, and David MJS Bowman. Increasing frequency and intensity of the most extreme wildfires on earth. *Nature ecology & evolution*, 8(8):1420–1425, 2024.

Francesca Di Giuseppe, Claudia Vitolo, Blazej Krzeminski, Christopher Barnard, Pedro Maciel, and Jesús San-Miguel. Fire weather index: The skill provided by the european centre for medium-range weather forecasts ensemble prediction system. *Natural Hazards and Earth System Sciences*, 20(8):2365–2378, 2020. doi: 10.5194/nhess-20-2365-2020.

Francesca Di Giuseppe, Joe McNorton, Anna Lombardi, and Fredrik Wetterhall. Global data-driven prediction of fire activity. *Nature Communications*, 16(1):2918, April 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-58097-7.

Beatriz Fuster, Jorge Sánchez-Zapero, Fernando Camacho, Vicente García-Santos, Aleixandre Verger, Roselyne Lacaze, Marie Weiss, Frederic Baret, and Bruno Smets. Quality assessment of proba-v lai, fapar and fcover collection 300 m products of copernicus global land service. *Remote Sensing*, 12(6), 2020. ISSN 2072-4292. doi: 10.3390/rs12061017. URL https://www.mdpi.com/2072-4292/12/6/1017.

Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Yuyang (Bernie) Wang, Mu Li, and Dit-Yan Yeung. Earthformer: Exploring space-time transformers for earth system forecasting. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 25390–25403. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/a2affd71d15e8fedffe18d0219f4837a-Paper-Conference.pdf.

Sebastian Gerard, Yu Zhao, and Josephine Sullivan. Wildfirespreadts: A dataset of multi-modal time series for wildfire spread prediction. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 74515–74529. Curran Associates, Inc., 2023. doi: 10.52202/075280-3258. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ebd545176bdaa9cd5d45954947bd74b7-Paper-Datasets_and_Benchmarks.pdf.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: https://doi.org/10.1002/qj.3803. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803.

Fantine Huot, R. Lily Hu, Nita Goyal, Tharun Sankar, Matthias Ihme, and Yi-Fan Chen. Next day wildfire spread: A machine learning dataset to predict wildfire spreading from remote-sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022. doi: 10.1109/TGRS.2022.3192974.

Jarosław Jasiewicz and Tomasz F. Stepinski. Geomorphons — a pattern recognition approach to classification and mapping of landforms. *Geomorphology*, 182:147–156, 2013. ISSN 0169-555X. doi: https://doi.org/10.1016/j.geomorph.2012.11.005. URL https://www.sciencedirect.com/science/article/pii/S0169555X12005028.

W. Matt Jolly, Mark A. Cochrane, Patrick H. Freeborn, Zachary A. Holden, Timothy J. Brown, Grant J. Williamson, and David M. J. S. Bowman. Climate-induced variations in global wildfire danger from 1979 to 2013. *Nature Communications*, 6:7537, 2015. doi: 10.1038/ncomms8537.

Jonathan Koh, François Pimont, Jean-Luc Dupuy, and Thomas Opitz. Spatiotemporal wildfire modeling through point processes with moderate and extreme marks. *The Annals of Applied Statistics*, 17(1):560 – 582, 2023. doi: 10.1214/22-AOAS1642. URL https://doi.org/10.1214/22-AOAS1642.

Spyridon Kondylatos, Ioannis Prapas, Gustau Camps-Valls, and Ioannis Papoutsis. Mesogeos: A multi-purpose dataset for data-driven wildfire modeling in the mediterranean. *Advances in Neural Information Processing Systems*, 36:50661–50676, 2023.

Spyros Kondylatos, Ioannis Prapas, Michele Ronco, Ioannis Papoutsis, Gustau Camps-Valls, María Piles, Miguel-Ángel Fernández-Torres, and Nuno Carvalhais. Wildfire danger prediction and understanding with deep learning. *Geophysical Research Letters*, 49(17):e2022GL099368, 2022. doi: https://doi.org/10.1029/2022GL099368. e2022GL099368 2022GL099368.

Spyros Kondylatos, Nikolas Papadopoulos, Gustau Camps-Valls, and Ioannis Papoutsis. Uncertainty-aware deep learning for wildfire danger forecasting. *Machine Learning: Earth*, 2(2):025008, jul 2026. doi: 10.1088/3049-4753/ae89b9. URL https://doi.org/10.1088/3049-4753/ae89b9.

Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 52, 1955. URL https://api.semanticscholar.org/CorpusID:9426884.

Saad Lahrichi, Jake Bova, Jesse Johnson, and Jordan Malof. Improved wildfire spread prediction with time-series data and the wsts+ benchmark. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2890–2900, 2026. doi: 10.1109/WACV61042.2026.00283.

Bencheng Liao, Shaoyu Chen, Xinggang Wang, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Chang Huang. Maptr: Structured modeling and learning for online vectorized hd map construction. (arXiv:2208.14437), January 2023. doi: 10.48550/arXiv.2208.14437. URL http://arxiv.org/abs/2208.14437. arXiv:2208.14437 [cs].

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2020. doi: 10.1109/TPAMI.2018.2858826.

Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. DAB-DETR: Dynamic anchor boxes are better queries for DETR. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=oMI9PjOb9Jl.

Shilong Liu, Tianhe Ren, Jiayu Chen, Zhaoyang Zeng, Hao Zhang, Feng Li, Hongyang Li, Jun Huang, Hang Su, Jun Zhu, and Lei Zhang. Detection transformer with stable matching. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6468–6477, 2023. doi: 10.1109/ICCV51070.2023.00597.

J. R. McNorton, F. Di Giuseppe, E. Pinnington, M. Chantry, and C. Barnard. A global probability-of-fire (pof) forecast. *Geophysical Research Letters*, 51(12):e2023GL107929, 2024. doi: https://doi.org/10.1029/2023GL107929. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2023GL107929. e2023GL107929 2023GL107929.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K. Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 25904–25938. PMLR, 2023. URL https://proceedings.mlr.press/v202/nguyen23a.html.

A.D. Nobre, L.A. Cuartas, M. Hodnett, C.D. Rennó, G. Rodrigues, A. Silveira, M. Waterloo, and S. Saleska. Height above the nearest drainage – a hydrologically relevant new terrain model. *Journal of Hydrology*, 404(1):13–29, 2011. ISSN 0022-1694. doi: https://doi.org/10.1016/j.jhydrol.2011.03.051. URL https://www.sciencedirect.com/science/article/pii/S0022169411002599.

Kemal Oksuz, Baris Can Cam, Sinan Kalkan, and Emre Akbas. Imbalance problems in object detection: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3388–3415, oct 2021. doi: 10.1109/TPAMI.2020.2981890.

Thomas Opitz, Florent Bonneu, and Edith Gabriel. Point-process based bayesian modeling of space–time structures of forest fire occurrences in mediterranean france. *Spatial Statistics*, 40:100429, 2020. ISSN 2211-6753. doi: https://doi.org/10.1016/j.spasta.2020.100429. URL https://www.sciencedirect.com/science/article/pii/S2211675320300233. Space-Time Modeling of Rare Events and Environmental Risks: METMA Conference.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: visual reasoning with a general conditioning layer. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Moritz Rösch, Michael Nolde, Tobias Ullmann, and Torsten Riedlinger. Data-driven wildfire spread modeling of european wildfires using a spatiotemporal graph neural network. *Fire*, 7(6), 2024. ISSN 2571-6255. doi: 10.3390/fire7060207. URL https://www.mdpi.com/2571-6255/7/6/207.

Wilfrid Schroeder, Patricia Oliva, Louis Giglio, and Ivan A. Csiszar. The New VIIRS 375m active fire detection data product: Algorithm description and initial assessment. *Remote Sensing of Environment*, 143:85–96, 2014. ISSN 0034-4257. doi: https://doi.org/10.1016/j.rse.2013.12.008. URL https://www.sciencedirect.com/science/article/pii/S0034425713004483.

Dmitrii Shadrin, Svetlana Illarionova, Fedor Gubanov, Ksenia Evteeva, Maksim Mironenko, Ivan Levchunets, Roman Belousov, and Evgeny Burnaev. Wildfire spreading prediction using multimodal data and deep neural network approach. *Scientific Reports*, 14(1):2606, January 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-52821-x.

Rackhun Son, Tobias Stacke, Veronika Gayler, Julia E. M. S. Nabel, Reiner Schnur, Lazaro Alonso, Christian Requena-Mesa, Alexander J. Winkler, Stijn Hantson, Sönke Zaehle, Ulrich Weber, and Nuno Carvalhais. Integration of a deep-learning-based fire model into a global land surface model. *Journal of Advances in Modeling Earth Systems*, 16(1):e2023MS003710, 2024. doi: https://doi.org/10.1029/2023MS003710. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2023MS003710. e2023MS003710 2023MS003710.

Qingyu Song, Changan Wang, Zhengkai Jiang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yang Wu. Rethinking counting and localization in crowds: A purely point-based framework. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3365–3374, 2021.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. (arXiv:1505.00387), November 2015. doi: 10.48550/arXiv.1505.00387. URL http://arxiv.org/abs/1505.00387. arXiv:1505.00387 [cs].

Mingxing Tan, Ruoming Pang, and Quoc V. Le. Efficientdet: Scalable and efficient object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10778–10787, Seattle, WA, USA, 2020. IEEE. ISBN 978-1-7281-7168-5. doi: 10.1109/CVPR42600.2020.01079. URL https://ieeexplore.ieee.org/document/9156454/.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6450–6459, 2018. doi: 10.1109/CVPR.2018.00675. URL https://ieeexplore.ieee.org/document/8578773/.

Gabriel Tseng, Anthony Fuller, Marlena Reil, Henry Herzog, Patrick Beukema, Favyen Bastani, James R. Green, Evan Shelhamer, Hannah Kerner, and David Rolnick. Galileo: Learning global & local features of many remote sensing modalities. (arXiv:2502.09356), 2025. doi: 10.48550/arXiv.2502.09356. URL http://arxiv.org/abs/2502.09356. arXiv:2502.09356 [cs].

Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3349–3364, October 2021. ISSN 1939-3539. doi: 10.1109/TPAMI.2020.2983686.

Martin J. Wooster, Gareth J. Roberts, Louis Giglio, David P. Roy, Patrick H. Freeborn, Luigi Boschetti, Chris Justice, Charles Ichoku, Wilfrid Schroeder, Diane Davies, Alistair M. S. Smith, Alberto Setzer, Ivan Csiszar, Tercia Strydom, Philip Frost, Tianran Zhang, Weidong Xu, Mark C. de Jong, Joshua M. Johnston, Luke Ellison, Krishna Vadrevu, Aaron M. Sparks, Hannah Nguyen, Jessica McCarty, Veerachai Tanpipat, Chris Schmidt, and Jesus San-Miguel-Ayanz. Satellite remote sensing of active fires: History and current status, applications and future requirements. *Remote Sensing of Environment*, 267:112694, 2021. doi: 10.1016/j.rse.2021.112694.

Chen Xu, Yao Xie, Daniel A. Zuniga Vazquez, Rui Yao, and Feng Qiu. Spatio-temporal wildfire prediction using multi-modal data. *IEEE Journal on Selected Areas in Information Theory*, 4: 302–313, 2023. doi: 10.1109/JSAIT.2023.3276054.

Yu Zhao, Sebastian Gerard, and Yifang Ban. Ts-satfire: A multi-task satellite image time-series dataset for wildfire detection and prediction. *Scientific Data*, 12(1):1817, November 2025a. ISSN 2052-4463. doi: 10.1038/s41597-025-06271-3.

Zeyang Zhao, Qilong Xue, Yuhang He, Yifan Bai, Xing Wei, and Yihong Gong. *Projecting Points to Axes: Oriented Object Detection via Point-Axis Representation*, volume 15086 of *Lecture Notes in Computer Science*, pp. 161–179. Springer Nature Switzerland, Cham, 2025b. ISBN 978-3-031-73389-5. doi: 10.1007/978-3-031-73390-1_10. URL https://link.springer.com/10.1007/978-3-031-73390-1_10.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable {detr}: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=gZ9hCDWe6ke.

A Feature List

Table 5: Input features used by WISP, grouped by model feature interface and native data resolution. All variables are harmonized to the VIIRS reference grid before entity extraction.

Group	Product	Temporal res.	Spatial res.	Count	Variables
Weather	ERA5 surface	Hourly	0.25°	4	t2m, skt, d2m, vpd
	ERA5 surface	Hourly	0.25°	3	wind_speed, blh, cape
	ERA5 surface	Hourly	0.25°	2	msl, sp
	ERA5 surface	Hourly accum.	0.25°	4	tp, lsp, cp, pev
	ERA5 surface	Hourly accum.	0.25°	5	ssr, ssrd, sshf, avg_snlwrf, avg_snswrf
	ERA5 surface	Hourly	0.25°	6	swvl1, swvl2, swvl3, swvl4, tcrw, tcrw_mask
	ERA5 pressure levels	Hourly	0.25°	10	temperature_700hpa, temperature_850hpa, relative_humidity_700hpa, wind_speed_300hpa, wind_speed_850hpa, vertical_velocity_700hpa, vertical_velocity_700hpa_mask, geopotential_700hpa, geopotential_850hpa, divergence_300hpa
Vegetation	Copernicus vegetation	10-day	300 m	4	gdmp, gdmp_mask, fapar, fcover
	VIIRS vegetation	8-day	500 m	1	lai
Land	Terrain	Static	0.1°	2	elevation, slope
	Hydrologic terrain	Static	90 m	1	hand
	Landform classes	Static	90 m	10	geomorphon_1, geomorphon_2, geomorphon_3, geomorphon_4, geomorphon_5, geomorphon_6, geomorphon_7, geomorphon_8, geomorphon_9, geomorphon_10
Human	Population	Annual	100 m	1	population_density
Fire	VIIRS active fire	12h	375 m	2	frp, active_fire
Total	–	–	–	55	–

B Historical Fire Feature Construction

WISP derives two additional fire-history features from the historical active fire observations:

$$\mathbf{A}^h = \left(\mathrm{AF}^h, \mathrm{FRP}^h\right), \quad \Psi_{\mathrm{fire}}\left(\mathrm{AF}^h\right) = \left(\mathrm{AF}^h_{\mathrm{mask}}, \mathrm{AF}^h_{\mathrm{rec}}\right), \tag{18}$$

where $\mathrm{AF}^h_{\mathrm{mask}}$ is the fire mask derived from historical active fire confidence, and $\mathrm{AF}^h_{\mathrm{rec}}$ is the normalized fire-recency feature. The raw fire-observation channels and the derived fire-history channels are then concatenated as

$$\mathbf{A}^h_{\mathrm{hist}} = \left[\mathbf{A}^h, \Psi_{\mathrm{fire}}\left(\mathrm{AF}^h\right)\right], \quad \mathbf{B}^h_{\mathrm{sum}} = \left(\mathbf{B}^h_{\mathrm{any}}, \mathbf{B}^h_{\mathrm{rec}}\right) \in \mathbb{R}^{2\times H\times W}. \tag{19}$$

Here, $\mathbf{B}^h_{\mathrm{any}}$ is a binary map indicating whether each pixel has any historical fire observation, and $\mathbf{B}^h_{\mathrm{rec}}$ records the normalized elapsed time since the most recent historical fire observation; pixels without historical fire are assigned a value of 1.

C Auxiliary Land and Fire Encoders

The auxiliary encoders produce four-level pyramids aligned with the historical context pyramid at resolutions $s \in \{128, 64, 32, 16\}$. The three encoders do not share parameters: Φ_{land} is an HRNet-style encoder (Wang et al., 2021) that keeps a 128×128 high-resolution branch active while constructing 64×64, 32×32, and 16×16 branches, followed by cross-scale exchange blocks; Φ_{af} encodes the full historical fire sequence $\mathbf{A}^h_{\mathrm{hist}}$. It combines temporal attention pooling over hourly fire features with a flattened-sequence convolutional path, thereby preserving both recent fire dynamics and aggregated historical fire evidence; Φ_{fire} is a lightweight convolutional pyramid for the two-channel fire-summary map $\mathbf{B}^h_{\mathrm{sum}}$. Although $\mathbf{B}^h_{\mathrm{sum}}$ has only two channels, it is lifted to the same multi-scale interface used by the other sources, which simplifies scale-wise condition injection. The subsequent gated fusion module regulates its contribution relative to other sources.

D Hungarian Matching Algorithm

In Table 6, the rows are predicted queries and columns are ground-truth fire clusters. Each entry is the matching cost $\mathcal{D}_{qk} = \lambda_{\mathrm{m,cls}}(-\hat{p}_q) + \lambda_{\mathrm{m,loc}}\|\hat{\mathbf{y}}_q - \mathbf{g}_k\|_1$. Bold entries indicate the selected one-to-one assignment.

Table 6: Illustration of the Hungarian query-to-cluster cost matrix.

	Target g_1	Target g_2	Target g_3
Query q_1	$\mathbf{-0.41}$	0.78	0.92
Query q_2	-0.20	0.10	0.65
Query q_3	0.55	$\mathbf{-0.35}$	0.50
Query q_4	0.48	0.44	$\mathbf{-0.05}$

Algorithm 1 Hungarian matching WISP

Input: Predicted query set $\hat{\mathcal{Y}} = \{(\mathbf{z}_q, \hat{\mathbf{y}}_q)\}_{q=1}^{Q}$; ground-truth fire-cluster set $\mathcal{G}_K = \{\mathbf{g}_k\}_{k=1}^{K}$; matching weights $\lambda_{\mathrm{m,cls}}, \lambda_{\mathrm{m,loc}}$; localization weight λ_{loc}; no-fire class weight w_{eos}.
Output: Set prediction loss $\mathcal{L}$.

Compute fire probabilities $\hat{p}_q \leftarrow \mathrm{softmax}(\mathbf{z}_q)_1$ for all $q = 1, \ldots, Q$.
if $K > 0$ **then**
  for $q = 1, \ldots, Q$ **do**
    for $k = 1, \ldots, K$ **do**
      Compute matching cost

$$\mathcal{D}_{qk} \leftarrow \lambda_{\mathrm{m,cls}}(-\hat{p}_q) + \lambda_{\mathrm{m,loc}} \left\| \hat{\mathbf{y}}_q - \mathbf{g}_k \right\|_1 .$$

    end for
  end for
  Solve the linear assignment problem

$$\pi^\star \leftarrow \mathrm{Hungarian}(\mathcal{D}),$$

  where $\pi^\star$ contains one-to-one query–cluster pairs.
else
  Set $\pi^\star \leftarrow \emptyset$.
end if
Initialize query labels $c_q^\star \leftarrow 0$ for all $q = 1, \ldots, Q$.
for all $(q, k) \in \pi^\star$ **do**
  Set $c_q^\star \leftarrow 1$.
end for
Compute classification loss over all queries:

$$\mathcal{L}_{\mathrm{cls}} \leftarrow \frac{1}{Q} \sum_{q=1}^{Q} \mathrm{CE}\left(\mathbf{z}_q, c_q^\star; \{w_{\mathrm{eos}}, 1\}\right).$$

if $|\pi^\star| > 0$ **then**
  Compute localization loss over matched queries:

$$\mathcal{L}_{\mathrm{loc}} \leftarrow \frac{1}{2|\pi^\star|} \sum_{(q,k)\in\pi^\star} \left\| \hat{\mathbf{y}}_q - \mathbf{g}_k \right\|_1 .$$

else
  Set $\mathcal{L}_{\mathrm{loc}} \leftarrow 0$.
end if
Combine losses:

$$\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{loc}} \mathcal{L}_{\mathrm{loc}}.$$

return $\mathcal{L}$.

E BENCHMARK AND DATA PREPROCESSING.

Preprocessing. We restrict candidate region cubes to the latitude band from $60°\mathrm{S}$ to $80°\mathrm{N}$, excluding polar regions. Then, VIIRS records are filtered to presumed vegetation fires, rounded to hourly time slots, and rasterized into two channels: an active fire confidence code and FRP. The active

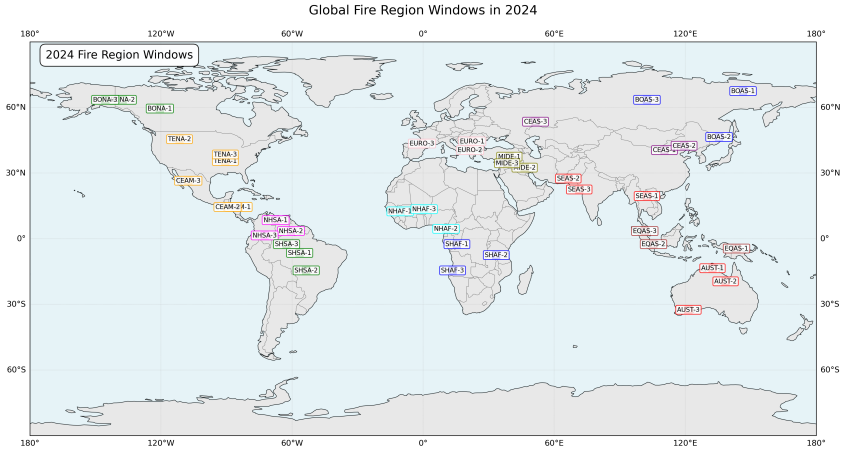


Fig. 3: Global distribution of the sampled fire-region windows in 2024. Three $3° \times 3°$ spatial windows are sampled from each of the 14 GFED regions. Labels indicate the corresponding GFED region and sampled window index.

fire convention is $-1 \rightarrow$ unobserved, $1/2/3 \rightarrow$ low/nominal/high confidence fire. When multiple records fall into the same pixel-hour, we keep only the highest confidence code. Then all variables are interpolated to a VIIRS $375\,\mathrm{m}$ resolution. Continuous meteorological and land variables are converted into physically comparable units, with transformations such as Kelvin-to-Celsius conversion, pressure conversion to hPa, $\log(1+x)$ for skewed variables, temporal differencing for accumulated fluxes, clipping for bounded land variables, and one-hot encoding for geomorphon classes. For variables whose valid physical support is sparse or sign-dependent, we additionally construct binary mask channels. Missing source values and provider-specific fill values are first standardized, and non-fire covariate gaps are filled with 0. The z-score normalization statistics used by the learning pipeline, mean and std, are estimated from the training dataset only.

Data augmentation. During training, WISP applies a random spatial jitter to the full 128×128 input entity before target construction. The model still receives the full context window, but supervision and evaluation are restricted to the central valid region $\texttt{valid_region_box} = [16, 16, 112, 112]$, corresponding to a 96×96 crop (see cyan box in Fig. 4). This design keeps contextual information around the target area while excluding boundary regions where shifted inputs may contain padding or incomplete neighbourhood context. It also prevents the model from exploiting a fixed sampling prior in which future fires always appear near the centre of the input window or some fixed locations (see Fig. 5). The sampled offset is recorded as $\texttt{spatial_jitter} = [dy, dx]$; for example, $[-4, 2]$ shifts the entity by 4 pixels upward and 2 pixels to the right in Fig. 4. The displayed $\texttt{subset_cover_xy} = 0.986$ is a structure-aware diagnostic used to select the best training snapshot: it measures how well a one-to-one subset of positive queries covers the ground-truth fire-cluster centres. Higher is better.

Training configuration. All model variants are trained with AdamW, learning rate 10^{-4}, and weight decay 10^{-4}. Training is distributed with Accelerate on 4 nodes with 4 A100 GPUs per node, for 16 GPUs in total. We use bfloat16 mixed precision and activation checkpointing. The per-GPU micro-batch size is 1; since $\texttt{grad_accum_steps} = 4$ by default, the effective batch size is $1 \times 16 \times 4 = 64$. The model uses dropout with probability 0.1 in the spatio-temporal attention backbone and query decoder. Each branch is trained for up to 5000 epochs, and the checkpoint with the best validation event mAP over AP@7, AP@14, and AP@21 is selected for test evaluation. The operating threshold for threshold-dependent query metrics is 0.5 unless otherwise specified.

Set-prediction hyperparameters. The default WISP configuration uses $Q = 10$ queries and location-only Hungarian matching, with $(\lambda_{\mathrm{m,cls}}, \lambda_{\mathrm{m,loc}}) = (0, 1)$. The query decoder uses two reference-refinement layers. The localisation loss weight is $\lambda_{\mathrm{loc}} = 5$, and the no-fire classification weight is $w_{\mathrm{eos}} = 0.1$. Classification supervision is retained even when the classification term is excluded from the matching cost. Appendix G compares alternative matching weights and query budgets of $Q = 10$ and $Q = 20$.

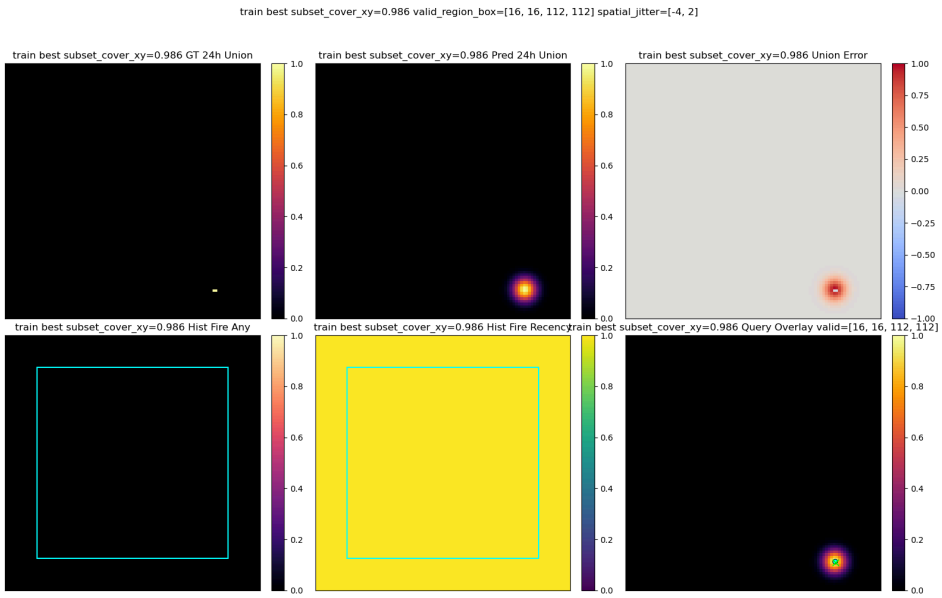


Fig. 4: The valid-region and spatial-jitter diagnostic during training process.

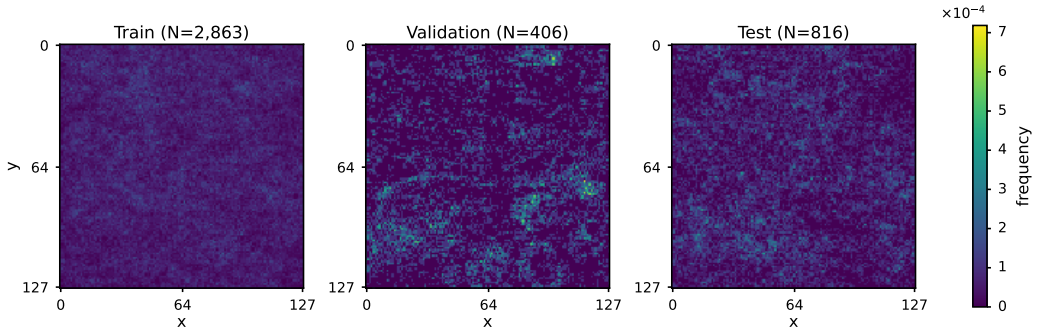


Fig. 5: Burning-pixel frequency maps for the train, validation, and test splits. For each split, the frequency at each 128×128 pixel is computed over the future 24-hour target windows. All panels use the same color scale, so spatial concentration and split-level sampling differences can be compared directly.

Future union and cluster targets. For supervision, future active fire observations are restricted to the central valid region. A 24-hour union fire mask is constructed by marking a pixel positive if it is a burning pixel at any future hour. Set targets are obtained by grouping the union mask into connected components using a 7×7 footprint, corresponding to a Chebyshev radius of 3 pixels. For each component, we compute an FRP-weighted centre; if the component has zero total FRP, we use the unweighted centroid. Components are ranked by total FRP mass, then by component size, and then by spatial position for deterministic tie-breaking. During training, only the first Q components are used as set targets; the full target set is retained for validation diagnostics and for reporting ground-truth truncation.

F Evaluation Metrics

WISP predicts a fixed set of query outputs $\hat{\mathcal{Y}} = \{(\hat{p}_q, \hat{\mathbf{y}}_q)\}_{q=1}^{Q}$, where $\hat{p}_q$ is the fire probability and $\hat{\mathbf{y}}_q \in [0,1]^2$ is a normalized point location. Ground truth is a set of fire-cluster centres $\mathcal{G} = \{(\mathbf{g}_k, m_k)\}_{k=1}^{K}$, where m_k is the FRP mass of cluster k. Distances are computed in pixel units on the valid VIIRS grid.

Event AP. For a spatial tolerance r, predictions are sorted by decreasing $\hat{p}_q$. A prediction is counted as a true positive if it lies within r pixels of an unmatched ground-truth centre; otherwise it is a false positive. This gives a precision–recall curve over all query predictions in the dataset, and AP@r is the area under this curve. We report AP@7, AP@14, and AP@21, and define

$$\mathrm{mAP} = \frac{1}{3}\left(\mathrm{AP@7} + \mathrm{AP@14} + \mathrm{AP@21}\right).$$

Mass coverage. At the operating threshold $\hat{p}_q \geq 0.5$, positive queries are matched one-to-one to ground-truth clusters by spatial similarity. Let π denote the matched query–cluster pairs, and let d_{qk} be the pixel distance between query q and cluster k. The FRP-mass coverage at radius r is

$$\mathrm{MassCov@}r = \frac{\sum_{(q,k)\in\pi} m_k \, \mathbb{I}[d_{qk} \leq r]}{\sum_{j=1}^{|\pi|} m_{(j)}},$$

where $m_{(j)}$ are the ground-truth cluster masses sorted in decreasing order. This denominator caps the metric by the number of active predictions, so the score measures whether the predicted query set covers the most important available fire-cluster mass.

Hit rate. Hit@r is the fraction of matched query–cluster pairs that fall within radius r:

$$\mathrm{Hit@}r = \frac{1}{|\pi|} \sum_{(q,k)\in\pi} \mathbb{I}[d_{qk} \leq r].$$

We use Hit@14 as the primary matched-query localization diagnostic.

Cluster recall denotes the target set retained for set matching, ranked by cluster FRP mass as described in Appendix E.

$$\mathcal{G}^{(Q)} = \{\mathbf{g}_k\}_{k=1}^{K_Q}, \qquad K_Q = \min(K, Q),$$

Let π^{set} denote the one-to-one query–target assignment and let $\eta = 0.5$ be the operating threshold. Cluster recall at spatial tolerance r is

$$\mathrm{ClusRec@}r = \frac{1}{K_Q} \sum_{(q,k)\in\pi^{\mathrm{set}}} \mathbb{I}[\hat{p}_q \geq \eta] \, \mathbb{I}[d_{qk} \leq r].$$

Thus, ClusRec@r measures the fraction of retained target clusters recovered by active queries within radius r. Entities without target clusters do not contribute to this recall metric.

Top-10 strong-cluster recall. For each entity, let

$$\mathcal{Q}_{10} = \mathrm{Top}_{\min(10,Q)}\{\hat{p}_q\}_{q=1}^{Q}$$

denote the indices of the highest-scoring predicted queries, and let

$$\mathcal{G}_{10} = \mathrm{Top}_{\min(10,K)}\{m_k\}_{k=1}^{K}$$

denote the strongest ground-truth clusters ranked by FRP mass. We compute a one-to-one minimum-distance assignment π_{10} between these two subsets. Top10Rec@r is then

$$\mathrm{Top10Rec@}r = \frac{1}{|\mathcal{G}_{10}|} \sum_{(q,k)\in\pi_{10}} \mathbb{I}[d_{qk} \leq r].$$

Unlike ClusRec@r, Top10Rec@r does not apply an operating score threshold: the prediction set is defined only by the ten highest-scoring queries. The reported value therefore measures how well the fixed top-10 prediction budget recovers the strongest observed fire clusters.

G Query Budget and Matching Cost

Table 7 reports five WISP configurations. Variants v1–v4 compare matching costs at $Q = 10$, whereas v3 and v5 compare query budgets under the same matching weights. The localisation loss weight is fixed at $\lambda_{\mathrm{loc}} = 5$ throughout. Variant v1 corresponds to WISP in the main comparison. Location-only matching removes the classification term from the assignment cost, but retains classification supervision in the training loss.

H Query Evolution During Training

Table 7: Query-budget and matching-cost ablations on the 816-entity test set. The five evaluation metrics are reported in %; higher is better. Red and blue denote the best and second-best results in each metric column.

Variant	Q	$\lambda_{\mathrm{m,cls}}$	$\lambda_{\mathrm{m,loc}}$	λ_{loc}	mAP	MassCov@14	Hit@14	ClusRec@14	Top10Rec@14
v1 (WISP)	10	0	1	5	**23.31**	**54.28**	**54.52**	**49.73**	**64.47**
v2	10	1	1	5	20.98	45.32	46.73	41.11	56.38
v3	10	1	2	5	23.10	46.35	48.82	43.75	61.50
v4	10	1	5	5	21.40	51.57	53.20	45.76	**61.76**
v5	20	1	2	5	**26.02**	**56.84**	**56.68**	**58.66**	59.28

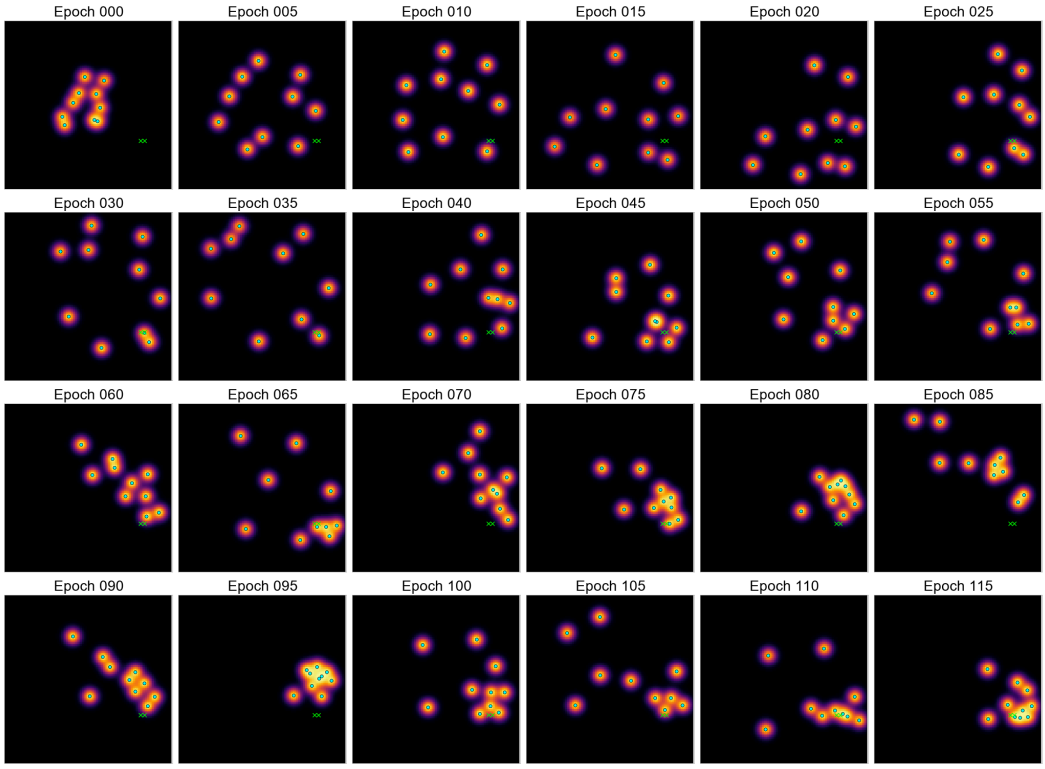


Fig. 6: Evolution of WISP query predictions on the first validation entity during the first 120 training epochs. The panels show the query overlay every five epochs, from epoch 0 to epoch 115. Cyan circles denote predicted fire queries and lime crosses denote ground-truth fire-cluster centres. Over training, the active query locations increasingly concentrate around the ground-truth future fire location. This visualization is used only as a qualitative diagnostic of training dynamics.

I ADDITIONAL QUALITATIVE RESULTS

In all appendix galleries, rows correspond to good, median, and failure cases from top to bottom, and columns show three examples within the same case group.

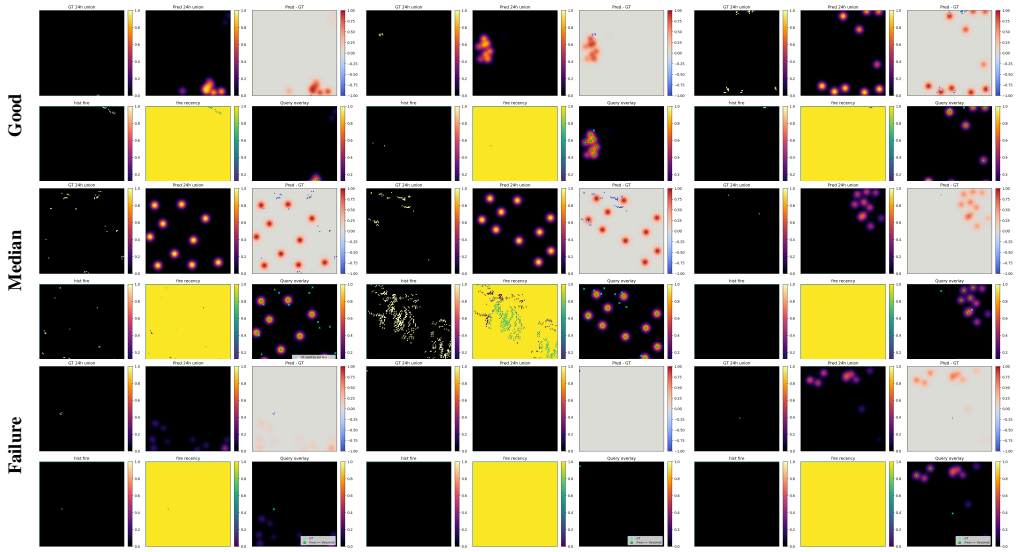


Fig. 7: Additional qualitative gallery for WISP variant v1.

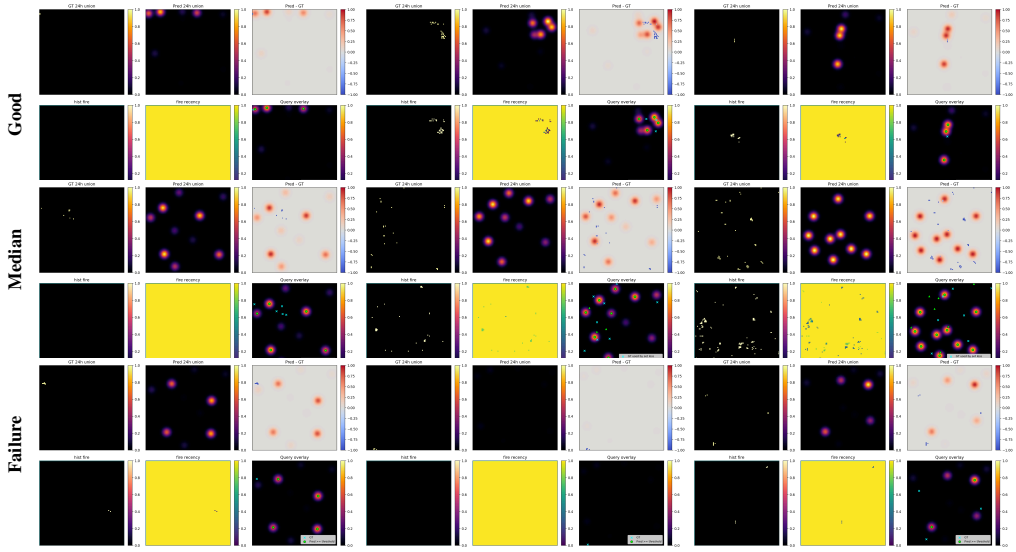


Fig. 8: Additional qualitative gallery for WISP variant v2.

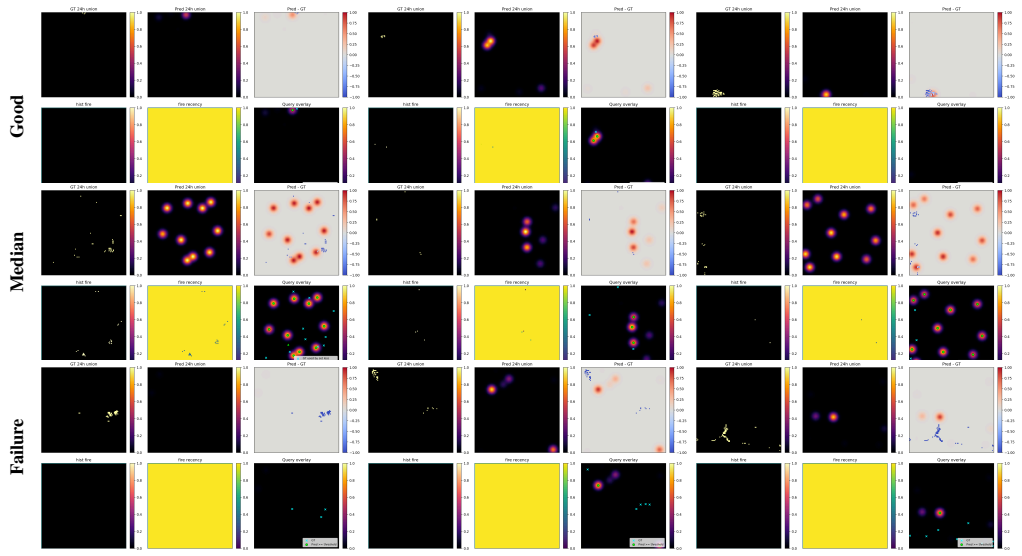


Fig. 9: Additional qualitative gallery for WISP variant v3.

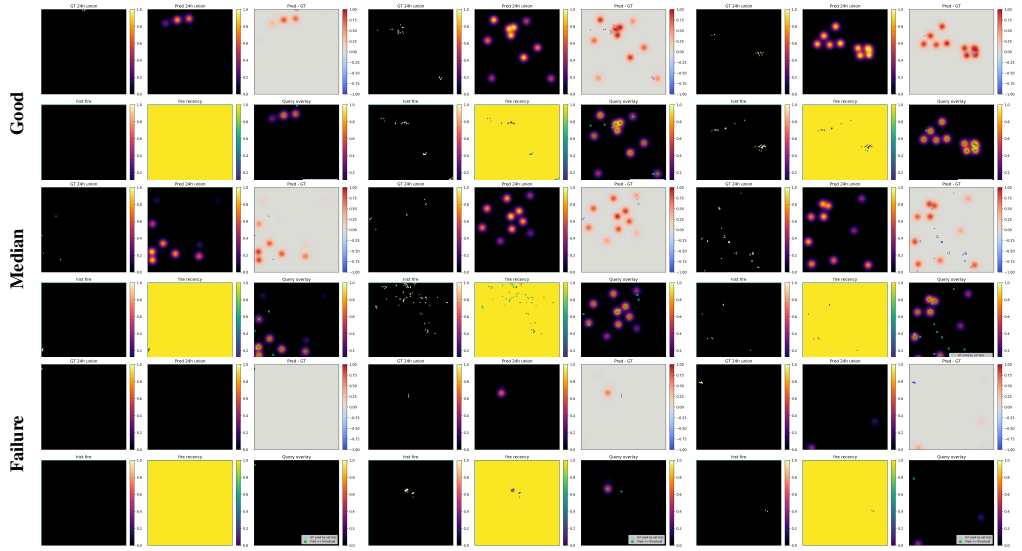


Fig. 10: Additional qualitative gallery for WISP variant v4.

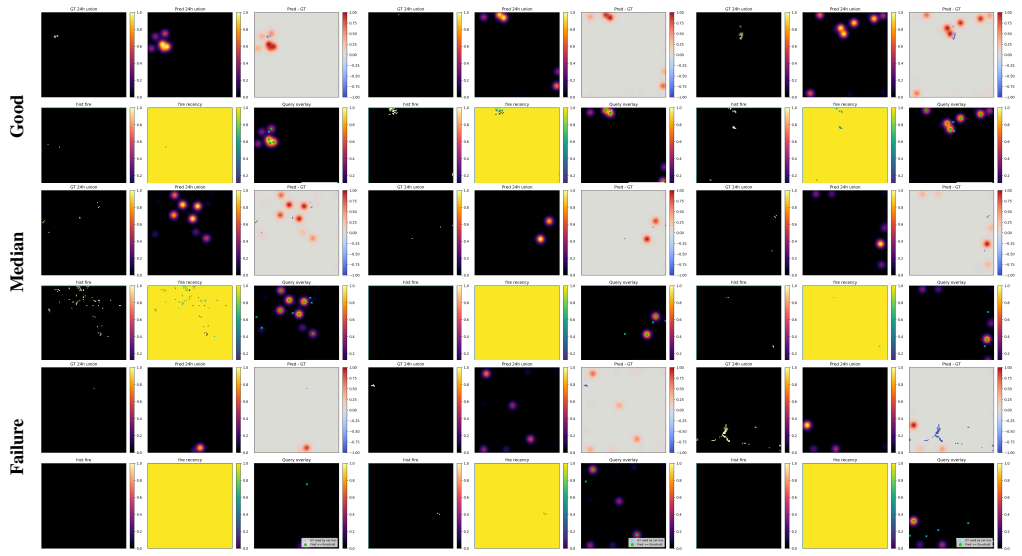


Fig. 11: Additional qualitative gallery for WISP variant v5.

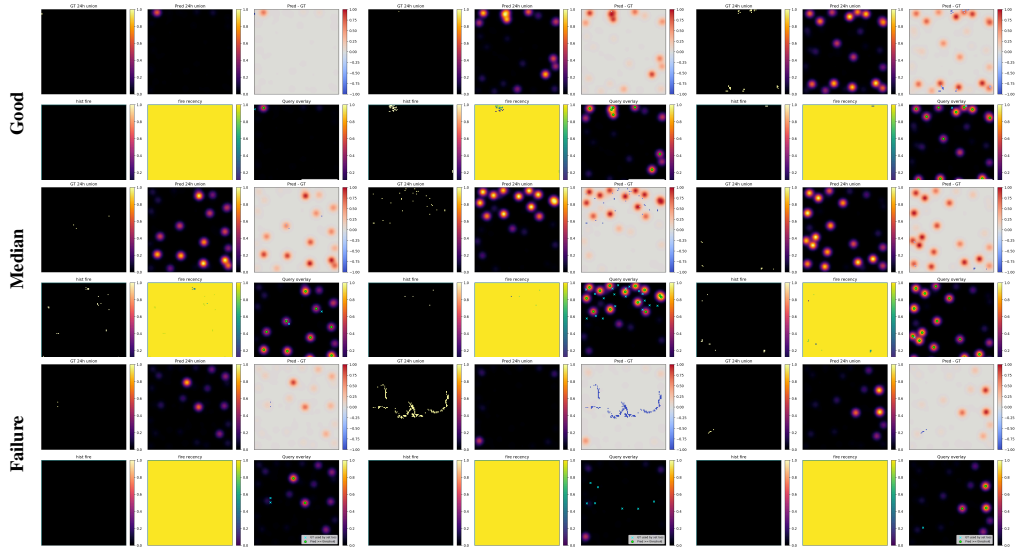


Fig. 12: Additional qualitative gallery for WISP variant v6.