

FireScope: Wildfire Risk Prediction with a Chain-of-Thought Oracle

Supplementary Material

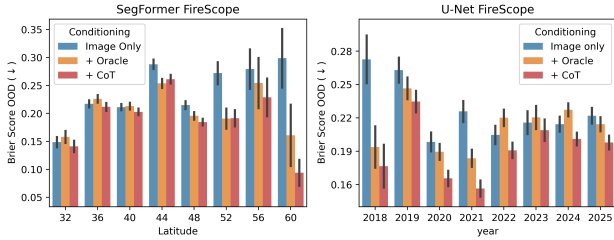


Figure 1. Error distribution of FireScope in Europe across latitudes and years. FireScope’s improvement is proportional to the distance from the distribution of training data, reaching its maximum in the 60th latitude and the year 2018. This suggests that FireScope is robust in varying operational settings.

Table 1. Model performance ID (MTBS fires in USA) and OOD (EFFIS fires in Europe). Top-3 are underscored. Top-1 is in bold.

Method		OOD: ROC AUC \uparrow		ID: ROC AUC \uparrow	
Condition	Encoder	events	pixels	events	pixels
Image only	AlphaEarth	0.692	0.632	0.782	0.633
	SegFormer	0.705	0.651	0.764	<u>0.613</u>
	U-Net	0.679	0.587	0.749	0.594
+ Climate	AlphaEarth	0.689	0.629	<u>0.778</u>	0.573
	SegFormer	0.557	0.518	0.728	0.571
	U-Net	0.591	0.559	0.750	0.575
+ Oracle	AlphaEarth	0.696	<u>0.653</u>	0.782	0.597
	SegFormer	0.713	0.642	0.730	0.588
	U-Net	0.698	<u>0.655</u>	0.743	0.610
+ CoT (FireScope)	AlphaEarth	0.717	<u>0.653</u>	0.766	0.596
	SegFormer	<u>0.727</u>	0.658	0.760	0.602
	U-Net	0.750	0.652	0.758	<u>0.615</u>

1. Wildfire events in the USA. Direct ID vs OOD comparison.

For completeness and direct comparison between OOD and ID performance, we also include 2284 real wildfire events and 1522 control areas in the USA from MTBS [16], filtering out controlled fires. They are included in FireScope-Bench. Evaluation results are presented in Tab. 1. This also allows for a direct comparison between ID and OOD performance. Comparing the difference in ROC AUC between FireScope and baseline, averaged across encoders, yields **-0.001 ID loss** against **+0.037 OOD gain** for wildfire events, and **-0.012 ID loss** against **+0.026 OOD gain** for pixels. Quantitatively, this demonstrates that FireScope’s OOD gain significantly outweighs ID loss.

2. Detailed Metrics

In-distribution (ID). As we have ground truth continuous risk rasters in the US, we use three metrics for evaluation:

Mean Squared Error (MSE) to quantify per-pixel error:

$$\text{MSE} = \frac{1}{N} \sum_i (x_i - y_i)^2 \quad (1)$$

Mean Absolute Error (MAE) to quantify per-pixel prediction error:

$$\text{MAE} = \frac{1}{N} \sum_i |x_i - y_i| \quad (2)$$

Structural Similarity Index (SSIM) [18] to assess spatial and structural consistency between predicted and reference risk maps:

$$\text{SSIM} = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (3)$$

where $C_1 = 10^{-4}$, $C_2 = 9 \cdot 10^{-4}$, and μ_x , σ_x , σ_{xy} are local mean, standard deviation and covariance, calculated in a gaussian sliding window with size of 11 pixels.

Out-of-Distribution (OOD). Since we only have wildfire events in Europe, we use probabilistic and spatial metrics to assess OOD performance. Risk is a combination between the probability of burning and expected intensity, which is directly correlated to the probability of observing a wildfire in a satellite image – places more likely to burn with greater intensity are consequently more likely to have observable fires. We therefore treat FireScope’s predicted risk scores as probabilities of observable wildfire events.

Brier score [2] is equivalent to MSE on the binary labels y :

$$\text{Brier Score} = \frac{1}{N} \sum_i (p_i - y_i)^2 \quad (4)$$

The *Area Under the Receiver Operating Characteristic Curve (ROC AUC)* [5] evaluates the model’s ability to discriminate between positive and negative samples:

$$\text{ROC AUC} = \int_0^1 \text{TPR}(\text{FPR}^{-1}(x)) dx = P(X_1 > X_0) \quad (5)$$

where TPR is true-positive rate, FPR is false-positive rate, X_1 is score for positive labels and X_0 is background score.

Expected Calibration Error (ECE) [10] quantifies the alignment between predicted probabilities and observed frequencies.

$$\text{ECE} = \sum_{b \in \text{bins}} \frac{n_b}{N} |\text{acc}_b - \text{conf}_b| \quad (6)$$

where bins are 15 equally spaced bins over $[0, 1]$, conf_b is the mean of predicted probabilities inside the bin, and acc_b is the accuracy of predictions whose predicted probabilities are inside the bin. N is the total number of pixels, while n_b is number of pixels inside bin b .

Finally, *Intersection over Union (IoU)* [7] measures the spatial overlap between predicted and reference burned areas:

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \quad (7)$$

where TP is true-positive, FP is false-positive, FN is false-negative.

Note that due to the stochastic nature of ignition, we treat non-burnt areas within wildfire tiles as background rather than negative samples [12]. Therefore, the maximum expected performance within individual images is limited, and predictions must be assessed based on the ability to distinguish wildfire events from controls areas in addition to quality of segmentation.

Oracles. To evaluate Oracles, we measure their classification performance on equivalent ordinal labels. We split risk level from $[0, 1]$ into ten evenly spaced bins $\{0, 1, \dots, 9\}$. For each training area, we take the corresponding ground truth risk raster and average the risk level across pixels. We thus map each area to an ordinal label from 0 to 9, representing the overall risk for that area. We then use the same metrics as above, but omit SSIM and IoU, as these are continuous predictions. We instead substitute them with *Quadratic Weighted Kappa (QWK)* [4] for ID data. QWK is a performance metric specifically designed for ordinal classification tasks.

Formally, given N items and K ordinal labels, let O be the observed $K \times K$ matrix of counts where O_{ij} denotes the number of instances rated as class i by the model and class j in the ground truth. Let E be the expected count matrix under the assumption of independent ratings, i.e.

$$E_{ij} = \frac{(\sum_k O_{ik})(\sum_k O_{kj})}{N}. \quad (8)$$

The QWK score is then defined as

$$\kappa = 1 - \frac{\sum_{i,j} (i-j)^2 O_{ij}}{\sum_{i,j} (i-j)^2 E_{ij}}. \quad (9)$$

Its values range from -1 (systematic disagreement) through 0 (chance-level agreement) up to 1 (perfect agreement). QWK measures **statistical agreement**, making it more sensitive to label distribution and chance-level predictions: for example, a naive model that always predicts the mean of the scale may achieve a deceptively high score in other metrics, but its QWK would remain close to zero, correctly reflecting the lack of real agreement with the ground truth.

3. Experiments Configurations

3.1. Models

Oracles. We select Qwen2.5-VL-7B-Instruct [1] as our Oracle for its strong open-source performance across multi-modal reasoning benchmarks and efficiency to train and deploy at only 7B parameters. We train two versions of it: one with CoT reasoning as outlined in Section 4.1 of the main paper, and one trained with supervised fine-tuning without CoT. We evaluate the classification performance of three additional Oracles:

- *GPT-5* [11]: a powerful general-purpose VLM.
- *Fire Weather Index (FWI)* [17]: a well-established physics-based index that estimates fire danger based on meteorological variables. We average it across summers between 2016 and 2025.
- *Climate MLP*: training a small multilayer perceptron head to classify based solely on the climatology vector.

Encoders. To evaluate the effectiveness of FireScope, we integrate the state-of-the-art vision encoder – SegFormer [19], offering complementary strengths in visual representation. We use SegFormer MiT-B5 version. To integrate Oracle’s output, we attach a FiLM-conditioned transformer decoder, which decodes hierarchically with cross-attention and modulation at multiple scales, using a learned target grid and skip fusions normalized with Group-Norm, followed by a small convolutional head that upsamples to the final resolution. We keep the pre-trained encoders frozen and fine-tune the FiLM and decoder head parameters.

Second, we include in our comparison AlphaEarth [3], a satellite embedding dataset in 10m resolution, as a strong in-domain baseline. The embedding model was pretrained on a number of tasks: self-distillation, natural language supervision, supervised reconstruction of land cover use, vegetation features, and monthly-aggregated climate. As it uses only Sentinel satellite imagery, it falls under the FireScope-Bench workflow. For training and evaluation, we concatenate AlphaEarth embeddings with the conditioning features and apply a simple two-layer convolutional neural network to produce the risk predictions.

Last, we design a lightweight U-Net [13] as a baseline trained from scratch. It uses four encoder and three decoder stages built from convolutional and residual blocks. Each encoder stage downsamples by a factor of 2-3 using strided convolutions, while feature depth doubles at each level. The decoder mirrors this hierarchy with transposed convolutions and skip connections that concatenate encoder activations at matching resolutions, followed by residual refinement blocks. The final 3×3 convolution produces the prediction map, cropped to match the target size. FiLM conditioning is applied to every stage.

Conditionings. For each choice of Encoder-Decoder,

we train up to four versions:

- Baseline: a baseline trained only on satellite images.
- Climate: conditioning on the raw climate vector.
- Oracle: conditioning on Qwen fine-tuned without CoT.
- CoT Oracle: conditioning on Qwen fine-tuned with CoT.

Ablations. We further include several additional experiments for ablation purposes:

- Qwen-enc: adding a transformer decoder with a lightweight convolutional upsampler head directly to Qwen2.5-VL-7B-Instruct to generate the raster in one forward pass.
- U-Net*: training the U-Net on the full volume of training data with translational augmentations (see Section 3 of the main paper).

We test each model ID on the 3935 testing set areas, and OOD on all wildfire and control events in Europe (see Section 3 of the main paper).

3.2. CoT Oracle Fine-Tuning

Whilst the ordinal labels defined in Section 3.2 of the main paper can be used to perform supervised fine-tuning on the Oracle, this would restrict it to outputting a single scalar score. To allow it to explore intermediate reasoning steps and refine its estimate iteratively, we fine-tune it with reinforcement learning (RL), which is unhindered by the lack of gradients for arbitrary-length outputs. For that, we use group-relative policy optimisation (GRPO) [15] – a novel reinforcement learning algorithm which foregoes the need for a critic model, thus incurring much less overhead compared to other widely used RL algorithms [6, 14].

In it, a reward function R is defined that quantifies the performance of model outputs after they are complete. Given a policy model $\pi_{\theta_{\text{old}}}(\cdot | p)$ parametrized by parameters θ_{old} and sampled training input p , GRPO samples a group of outputs $o_1, \dots, o_n \sim \pi_{\theta_{\text{old}}}(\cdot | p)$ and estimates advantages \hat{A}_i by comparing individual rewards $r_i = R(o_i)$ to the group’s normalized mean:

$$\hat{A}_i = \frac{r_i - b}{\sigma}, \quad b = \frac{1}{n} \sum_{i=1}^n r_i, \quad \sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (r_i - b)^2} \quad (10)$$

Then, the following objective is maximized:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{p \sim \mathcal{X}, \{o_i\}_{i=1}^n \sim \pi_{\theta_{\text{old}}}(\cdot | p)} \left[\frac{1}{n} \sum_{i=1}^n \min(d_i \hat{A}_i, \text{clip}(d_i, 1 - \varepsilon, 1 + \varepsilon) \hat{A}_i) - \beta D_{\text{KL}}(\pi_{\theta} \| \pi_{\text{ref}}) \right] \quad (11)$$

where

$$d_i = \frac{\pi_{\theta}(o_i | p)}{\pi_{\theta_{\text{old}}}(o_i | p)} \quad (12)$$

$$D_{\text{KL}}(\pi_{\theta} \| \pi_{\text{ref}}) = \mathbb{E}_{o \sim \pi_{\theta}(\cdot | p)} \left[\log \frac{\pi_{\theta}(o | p)}{\pi_{\text{ref}}(o | p)} \right] \quad (13)$$

Here, π_{ref} is a frozen copy of the model before fine-tuning.

By maximizing this objective, we update the policy to increase the probability of outputs with positive advantages and reduce it for those with negative ones. D_{KL} is a regularizing term that penalizes updates that change the policy too radically.

Reward Design. Our reward function consists of three components: an accuracy term, a formatting term, and a label-frequency-based reweighting factor. The final reward is defined as

$$R = w(\hat{y}) (0.9 R_{\text{acc}} + 0.1 R_{\text{fmt}}), \quad (14)$$

where $w(\hat{y})$ is a class-dependent weight determined by the frequency of the ground-truth label \hat{y} in the training set.

Given a correctly parsed model prediction y and the corresponding ground-truth label \hat{y} , we compute the absolute error

$$d = |y - \hat{y}|. \quad (15)$$

The accuracy reward is defined as

$$R_{\text{acc}} = \begin{cases} 1.0, & \text{if } d = 0, \\ 0.5, & \text{if } d = 1, \\ 0.1, & \text{if } d = 2, \\ 0, & \text{otherwise.} \end{cases} \quad (16)$$

The formatting reward is

$$R_{\text{fmt}} = \begin{cases} 1, & \text{if the generated output is parseable,} \\ 0, & \text{otherwise.} \end{cases} \quad (17)$$

Let $f(c)$ denote the frequency of class c in the training set, and let

$$f_{\min} = \min_c f(c), \quad f_{\max} = \max_c f(c). \quad (18)$$

We assign each class c a weight by linearly interpolating between higher weights for rarer classes and lower weights for more frequent classes:

$$w(c) = 1 + \left(\frac{f_{\max}}{f_{\min}} - 1 \right) - \left(\frac{f_{\max}}{f_{\min}} - 1 \right) \frac{f(c) - f_{\min}}{f_{\max} - f_{\min}}. \quad (19)$$

Thus, rarer labels receive larger rewards, which compensates for class imbalance during training.

3.3. Training Setup.

We perform full model fine-tuning on the CoT Oracle using the GRPO framework with a learning rate of 1×10^{-5} , four generations per prompt, and mixed-precision (bfloat16) training. We set per-device batch size to 16. The output length is capped at 1024 tokens, and optimization uses $D_{\text{KL}} = 0.01$. We train for 45 epochs over two days on four NVIDIA H200 GPUs.

We fine-tune the Oracle without CoT using cross-entropy loss on the next-token prediction of the target risk digit. Training runs with mixed-precision (bfloat16), using a batch size of 16, and a learning rate of 1×10^{-5} , and lasts less than 24 hours on four NVIDIA H200 GPUs.

For the VLM-encoder, we train Qwen2.5-VL-7B-Instruct with mixed-precision (bfloat16), a batch size of 16, and a learning rate of 1×10^{-5} , using the same training objective as for the Encoder-Decoder, as described in Section 4.2 of the main paper .

All Encoder-Decoder architectures are optimized with the AdamW optimizer [8] ($\beta_1=0.9, \beta_2=0.95$, weight decay = 10^{-5}) and a cosine annealing learning rate schedule [9], applying gradient clipping at a norm of 1.0. We use a per-device batch size of 32. For each, we customize the learning rate and total training epochs:

- SegFormer: lr = 1×10^{-5} , epochs = 500;
- U-Net: lr = 1×10^{-3} , epochs = 1000;
- AlphaEarth: lr = 1×10^{-3} , epochs = 1000.

To isolate the effects of our ablations, we keep those consistent for each Encoder-Decoder across experiments. Each training lasts less than 24 hours on four NVIDIA H200 GPUs.

After each training, we benchmark the checkpoint with the best evaluation performance.

We acknowledge the importance of gauging the sensitivity of the various hyperparameters, but due to the extensive number of experiments already required and the significant computational cost associated with each training run, a full hyperparameter sensitivity analysis is beyond our practical budget. Thus, we rely on configurations grounded in common practices in the field.

4. Ablation results

Metrics of ablation methods are reported in Table 2

Table 2. Metrics of ablation methods.

Method Conditioning	Encoder	OOD wildfire events			OOD wildfire pixels		ID wildfire risk rasters		
		Brier ↓	ROC AUC ↑	ECE ↓	ROC AUC ↑	IoU@0.5 ↑	MSE ↓	SSIM ↑	MAE ↓
n/a	VLM-enc	0.200	0.738	0.069	0.647	0.176	0.050	0.468	0.182
Image only	Unet*	0.208	0.699	0.058	0.619	0.173	0.019	0.620	0.090
Climate Data	Unet*	0.272	0.598	0.171	0.573	0.153	0.018	0.622	0.081

5. General-Purpose Capabilities of CoT Oracle (FireScope)

After prompting CoT Oracle to provide its reasoning and overall risk score, it may be of interest to users to acquire answers to follow-up questions or continue conversing. For this reason, we perform simple quantitative and qualitative analyses of CoT Oracle’s general purpose capabilities. In this section, we refer to CoT Oracle as simply Oracle.

Quantitative analysis. To quantitatively assess how much of Oracle’s general-purpose capabilities have been sacrificed to specialize in wildfire risk prediction, we evaluate it on the multiple-choice subset of the Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark (MMMU) [20]. Oracle achieves an accuracy of **51.12%**. This is a drop of 7.48% from Qwen2.5-VL 7B’s performance before fine-tuning, which is not insignificant,

but CoT Oracle still attains a better accuracy than many other popular VLMs of comparable size. This suggests that even after fine-tuning, Oracle remains surprisingly competent across diverse multimodal tasks, retaining much of its original reasoning ability.

Qualitative analysis. To better understand Oracle’s abilities beyond producing scalar wildfire risk scores, we explore its behavior in extended, free-form interactions that mirror realistic user queries and decision-making scenarios. We provide an example conversation where the user asks follow-up questions about building a house in the area after attaining Oracle’s wildfire risk estimate. The area is at $48.7392^\circ N, -113.4151^\circ W$, the Saint Mary lakes in Montana.



Example: Follow-up questions about the placement of a house in the area.

Oracle: To determine the wildfire risk in the area based on the satellite image and climate data, we need to analyze the following factors:

1. **Vegetation Density:** The satellite image shows areas with varying shades of green, indicating different levels of vegetation density.

Dense vegetation can increase fire risk due to more fuel available for fires to spread.

2. **Climate Data:**

- **Average Temperature:** Lower temperatures generally reduce fire risk because fires require higher temperatures to ignite and sustain. However, the average temperatures in the image region are relatively low, suggesting that fire risk might be moderate but not extremely high.

- **Average Humidity:** Higher humidity reduces fire risk by providing moisture that can help extinguish fires. The climate data shows lower humidity in some months, which could increase fire risk.

- **Average Wind Speed:** Stronger winds can increase fire risk by spreading flames faster. The climate data indicates variable wind speeds, with some months showing higher wind speeds, which could exacerbate fire spread.

- **Average Wind Direction:** The wind direction is mostly from the west or northwest, which could align with potential slope directions in the area, increasing fire risk.

3. **Slope:** The satellite image shows areas with varying elevations, which could indicate slopes. Slopes can increase fire risk as they allow fires to spread more easily uphill.

Analysis:

- **Vegetation Density:** The image shows a mix of dense and less dense vegetation. Dense vegetation increases fire risk.

- **Climate Data:**

- **Temperature:** Low temperatures suggest moderate fire risk.

- **Humidity:** Lower humidity in some months suggests increased fire risk.

- **Wind Speed:** Variable wind speeds, with some higher values, suggest increased fire risk.

- **Wind Direction:** The wind direction is mostly from the west or northwest, which could align with potential slope directions, increasing fire risk.

- **Slope:** The image shows areas with varying elevations, which could indicate slopes. Slopes increase fire risk.

Risk Assessment:

- **Vegetation Density**: Moderate risk.
- **Climate Data**:
- **Temperature**: Moderate risk.
- **Humidity**: High risk.
- **Wind Speed**: High risk.
- **Wind Direction**: High risk.
- **Slope**: High risk.

Final Classification:

The overall risk assessment suggests a high level of wildfire risk in the area. The combination of dense vegetation, dry conditions, strong winds, and potential slopes all contribute to a high risk of wildfires.

FINAL ANSWER:

8

User: If I had to build a house here, where would you recommend I build it and why? Give me a specific area in the image.

Oracle: Given the high wildfire risk in the area, it's crucial to choose a location that minimizes exposure to potential fire hazards. Here are some recommendations based on the satellite image and the analysis of the wildfire risk factors:

1. **Avoid Areas Near Vegetation**: Stay away from areas with dense vegetation, as these are more likely to catch fire and spread quickly.
2. **Consider Elevation**: Higher elevation areas may have better air circulation, which can help dissipate heat and reduce the risk of fire.
3. **Avoid Low-Lying Areas**: Low-lying areas can accumulate heat and moisture, making them more susceptible to fire.

Recommended Area:

Based on the satellite image, a good recommendation would be to build the house in the upper part of the image, specifically in the lighter green areas that appear to be less dense with vegetation. These areas are likely to have better air circulation and reduced fuel for fires.

Specific Recommendation:

- **Upper Left Quadrant**: This area appears to have less dense vegetation and is situated at a higher elevation, which could provide better protection against fire.

Reasoning:

- **Vegetation Density**: The upper left quadrant has lighter green areas, indicating less dense vegetation, which reduces the amount of fuel available for fires.
- **Elevation**: The higher elevation in this area suggests better air circulation, which can help dissipate heat and reduce the risk of fire.
- **Proximity to Water**: There is a body of water nearby, which can act as a natural barrier against fire spread.

By building in the upper left quadrant, you minimize the risk of your home being directly affected by a wildfire while also benefiting from the natural barriers provided by the water and the elevated terrain.

Final Answer:

Build the house in the upper left quadrant of the image, specifically in the lighter green areas that appear to be less dense with vegetation.

FINAL ANSWER:

Upper left quadrant, avoiding dense vegetation areas.

User: What other advantages or disadvantages are there to building my house there, unrelated to wildfires?

Oracle: Building your house in the upper left quadrant of the image, which appears to be in a less dense vegetation area and at a higher elevation, offers several advantages and disadvantages unrelated to wildfires. Here's a detailed breakdown:

Advantages:

1. **Better Air Circulation**: Higher elevation areas typically experience better air circulation, which can help cool the surrounding environment and reduce the risk of heat accumulation. This can make the area more comfortable year-round.
2. **Proximity to Water**: The presence of a body of water nearby can provide a natural buffer zone, reducing the risk of direct fire threats. Additionally, water can be used for firefighting purposes if needed.
3. **Scenic Views**: Higher elevation areas often offer panoramic views of the landscape, which can be visually appealing and provide a sense of tranquility.
4. **Reduced Noise Pollution**: Higher elevation areas are usually quieter than lower-lying areas, as they are farther from urban centers and industrial noise sources.
5. **Potential for Wildlife**: Depending on the local ecosystem, higher elevation areas might attract more wildlife, offering opportunities for nature observation and enjoyment.

Disadvantages:

1. **Cold Weather**: Higher elevation areas tend to experience colder temperatures during winter, which can be uncomfortable for residents who prefer warmer climates. This might require additional heating systems and insulation.
2. **Snow Accumulation**: Snowfall in higher elevation areas can be significant, leading to snow accumulation that might require regular maintenance and snow removal.
3. **Limited Infrastructure**: Higher elevation areas might lack certain infrastructure such as reliable electricity, water supply, and road access. This could pose challenges for daily living and emergency services.
4. **Increased Travel Time**: If the area is far from major cities or towns, travel time to essential services like hospitals, grocery stores, and entertainment venues might be longer.
5. **Potential for Snow and Ice Damage**: While snow can be beautiful, it can also cause damage to structures, especially if not properly maintained. Snow and ice can accumulate on roofs and windows, potentially leading to structural issues.

Conclusion:

The upper left quadrant of the image offers a balance of advantages and disadvantages. The benefits include better air circulation, proximity to water, scenic views, and reduced noise pollution. However, it also comes with challenges such as cold weather, snow accumulation, limited infrastructure, and potential for snow and ice damage. It's important to weigh these factors carefully before making a decision. If you are willing to accept the trade-offs and enjoy the unique bene-

fits of a higher elevation area, it could be a great choice for your new home.

6. Prompts used

In this section, we provide all prompts used for the various experiments related to the creation and evaluation of FireScope.

CoT Oracle prompt. This is the prompt provided to CoT Oracle to establish acceptable baseline performance. It is then fine-tuned with this prompt using GRPO. <climate> and <image> are placeholders for where we insert the climate description and the image token.

You are generating a wildfire risk raster for an area based on a satellite images and climate data.

Some factors which increase wildfire risk are: dense and dry vegetation, dry and hot climate, and strong winds, particularly if they align with slopes.

You must reason about the climate data and satellite image and explain in detail the risk level for all visible parts of the satellite image.

At the end, you must finish with a general classification for the wildfire risk in the area from 0 to 9. Your output must end with:

FINAL ANSWER:

n

Where n is a number between 0 and 9, on a newline.

Here is the climate data:

<climate>

Here is the satellite image:

<image>

Oracle without CoT prompt. This is the prompt provided to Oracle to establish acceptable baseline performance before supervised fine-tuning. It is then fine-tuned with this prompt minimizing next-token cross-entropy. <climate> and <image> are placeholders for where we insert the climate description and the image token.

You are evaluating the risk of wildfire in an area based on a satellite image and climate data. Some factors which increase wildfire risk are: dense and dry vegetation, dry and hot climate, and strong winds, particularly if they align with slopes.

You must respond with a single digit: the risk of wildfire in the area from 0 to 9, where 0 is lowest possible risk and 9 is highest.

Here is the climate data:

<climate>

Here is the satellite image:

<image>

Qwen+decoder prompt. This is the prompt provided to Qwen with a decoder head attached before training it to generate the risk raster in one forward pass (see Section 5.3 of the main paper). <climate> and <image> are placeholders for where we insert the climate description and the image token.

You are generating a wildfire risk raster for an area based on a satellite images and climate

data.

Some factors which increase wildfire risk are: dense and dry vegetation, dry and hot climate, and strong winds, particularly if they align with slopes.

You must consider all the data you are given and generate the risk raster for this area.

Here is the climate data:

<climate>

Here is the satellite image:

<image>

Perturb CoT prompt. This is the prompt provided to Qwen/Qwen3-4B-Instruct-2507 in order to generate perturbed CoTs to measure fidelity (see Section 5.4.2 of the main paper). <CoT> is a placeholder for where we insert Oracle's original CoT. We provide the prompt for when the new CoT should steer towards the highest risk score. The opposite case is analogical.

You will receive a chain of thought (CoT) by a wildfire expert evaluating the wildfire risk in the area.

TASK: Produce a new chain-of-thought (CoT) so it argues for the HIGHEST wildfire risk level without making up new facts.

HARD REQUIREMENTS:

- COMPLETELY REMOVE the mention of any factors that decrease risk.
- Emphasize risk-increasing factors already implied in the CoT.
- Only use area features already mentioned in the CoT.
- Do NOT explicitly mention the risk level in the new chain of thought. Use qualitative wording only.

INPUT:

CoT: <CoT>

OUTPUT: ONLY the new, rewritten CoT. NOTHING ELSE. Do NOT output the original CoT. Do NOT give your output an explanation or a title. Output ONLY new CoT and nothing else.

Paraphrase CoT prompt. This is the prompt provided to Qwen/Qwen3-4B-Instruct-2507 in order to generate paraphrased CoTs to measure consistency (see Section 5.4.2 of the main paper). <CoT> is a placeholder for where we insert Oracle's original CoT.

You are given a chain of thought from a wildfire risk model. Your task is to paraphrase it - rewrite it without changing its meaning.

Requirements:

- Rewrite the text by changing the wording slightly while keeping the structure, order, and meaning exactly the same.

- Keep the length and level of detail as close as possible to the original.

- Do not add, remove, or reorder any information - only substitute words or short phrases with near synonyms.

CoT:

<CoT>

Golden CoT generation prompt. This is the prompt provided to GPT-5 in order to generate golden CoTs for the expert study (see Section 5.4.1 in the main body). <risk>, <climate>, and <image> are placeholders for where we insert the correct risk score, the climate description, and the image token respectively.

You are generating a wildfire risk raster for an area based on a satellite images and climate data. Some factors which increase wildfire risk are: dense and dry vegetation, dry and hot climate, and strong winds, particularly if they align with slopes. You must reason about the climate data and satellite image and explain in detail the risk level for all visible parts of the satellite image.

At the end, you must finish with a general classification for the wildfire risk in the area from 0 to 9. Your output must end with:

FINAL ANSWER:

<risk>

You must start as if you don't know anything about the risk level, and as you go through the data, you analyze it and your chain of thought leads to the given final answer. DO NOT just repeat the data - any data you go through, you analyze in the context of wildfire risk. Your chain of thought eventually naturally arrives at the final answer.

You do not know the label in advance, and as you think and reason through the data, you come up with a final answer of <risk> on your own.

You must ALWAYS keep hidden that the final answer has been revealed to you in advance. You must arrive to it on your own, simulating forward reasoning.

CLIMATE CONDITIONS: <climate>

SATELLITE IMAGE:

<image>

Expert survey CoT preparation prompt. This is the prompt provided to Qwen/Qwen3-4B-Instruct-2507 in order to summarize the golden and FireScope's CoTs and omit conclusions before giving them to experts for the expert study in Section 5.4.1 of the main paper. <CoT> is a placeholder for where we provide the original CoT.

You are given a chain of thought from a wildfire risk model. Your task is to summarize the wildfire risk factors considered in the model's reasoning clearly and objectively, without specifying the risk level in any way.

Requirements:

- Do **NOT** use any words or terms which allude to a particular risk level (e.g. "moderate", "high", etc). This is critical.

- Do **NOT** include or infer the model's final

risk classification, numerical risk score, or probability.

- Summarize only the evidence, reasoning steps, and factors considered.

- Maintain a **completely neutral tone** so that the text does not suggest a final risk level **at all**.

- Keep the summary as short as possible, mention only the key points, and omit all redundant phrases.

- Output should read like a professional briefing note of all risk increasing or dampening factors in the area. It should **not** specify a particular risk level in any way.

Output format:

A single, well-structured explanation that captures the risk factors considered in the model's reasoning clearly, **without** alluding to any risk level **at all**, quantitatively or qualitatively.

Do not use any terms that imply a risk level, such as 'high,' 'moderate,' 'low,' 'severe,' 'elevated,' or any numerical score. Avoid any language that suggests the likelihood, intensity, or severity of a fire. Focus only on the evidence, factors, and their relationships - without concluding or ranking the risk.

<CoT>

CoT Oracle Evaluation Prompt on MMMU. This is the prompt provided to CoT Oracle to evaluate its general-purpose performance on the multiple-choice subset of the Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark (MMMU) [20]. For the results, refer to Section 5. <question> and <options> are placeholders for where we insert the questions and the answer choices.

You are an expert AI assistant solving multiple-choice questions.

You see a question, one or more images, and several answer options.

Carefully read the question and analyze the image(s).

Then, reason step by step and choose the single best option.

Question:

<question>

Options:

<options>

First, think through the problem. Then, on the last line, output:

FINAL ANSWER: X

where X is the letter (A, B, C, or D, etc.) of the correct option.

7. Error study

See figures 1, 2, 3, 4 for detailed error distribution study.

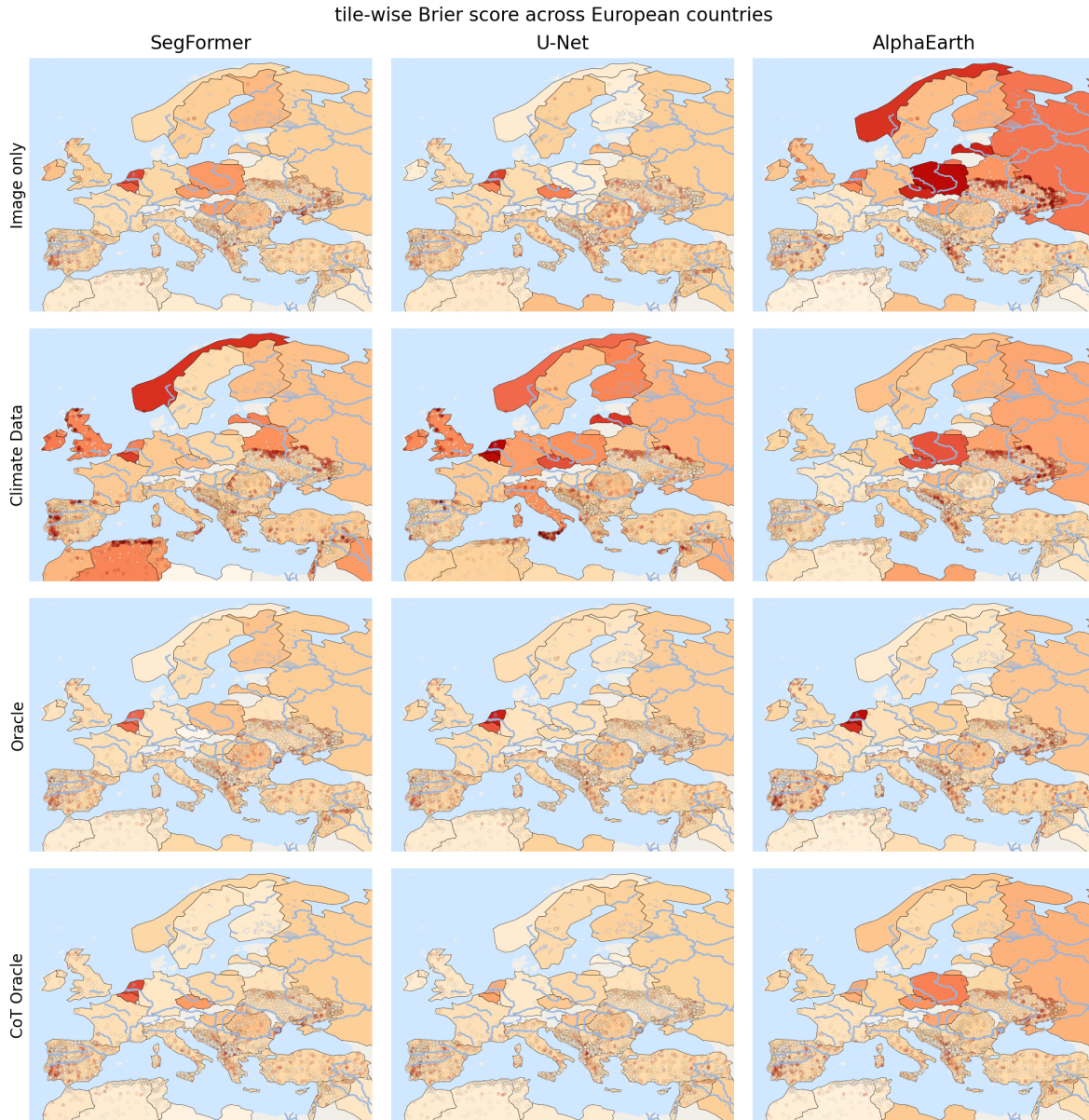


Figure 2. tile-wise Brier Score

8. Examples

In Figures 5, 6, 7, 8, we provide exemplary predictions of CoT Oracle.
 In Figure 9, we provide examples for high fidelity predictions of U-Net
 FireScope (see Section 5.4.2 of the main paper).

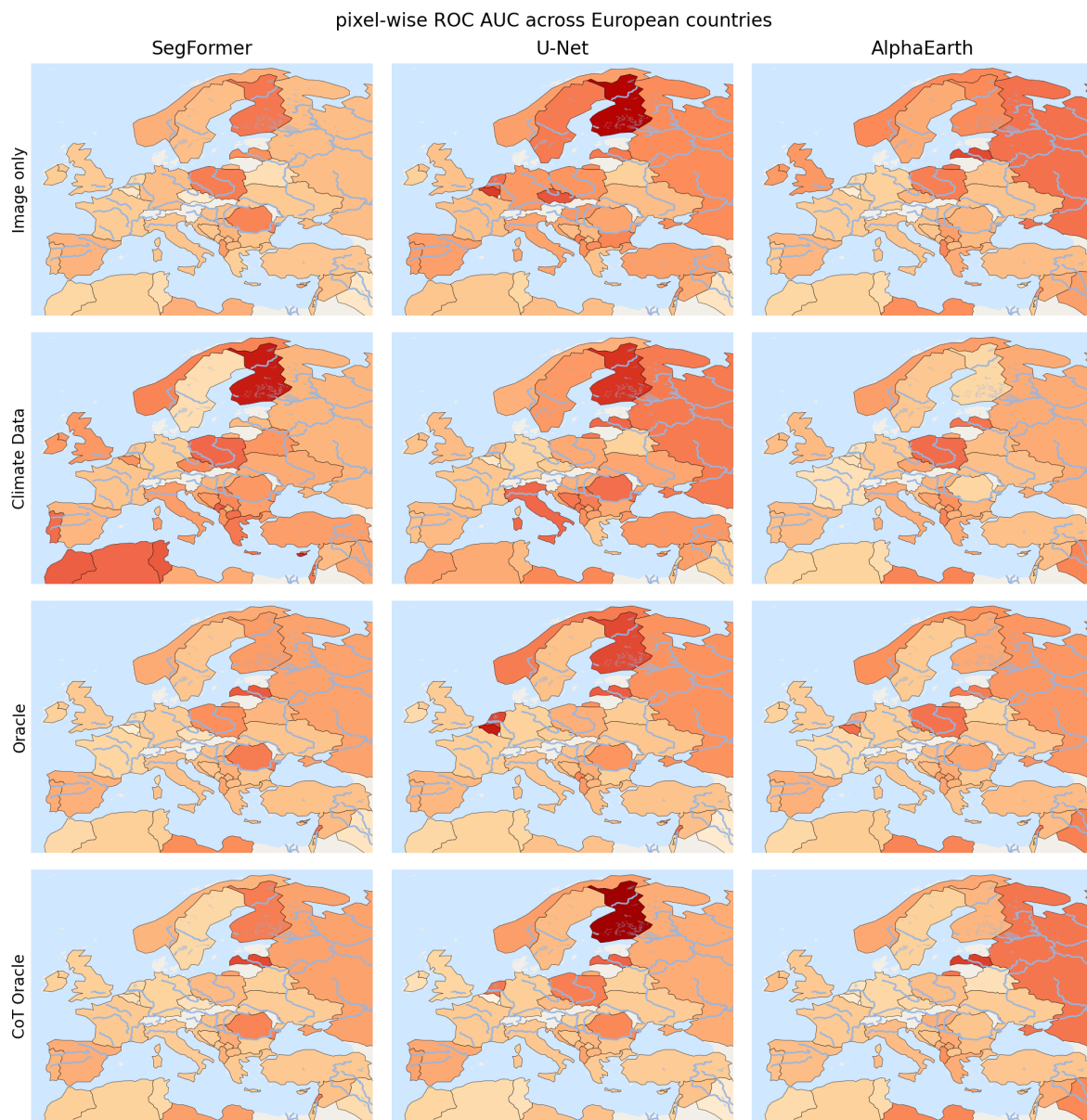


Figure 3. pixel-wise ROC AUC

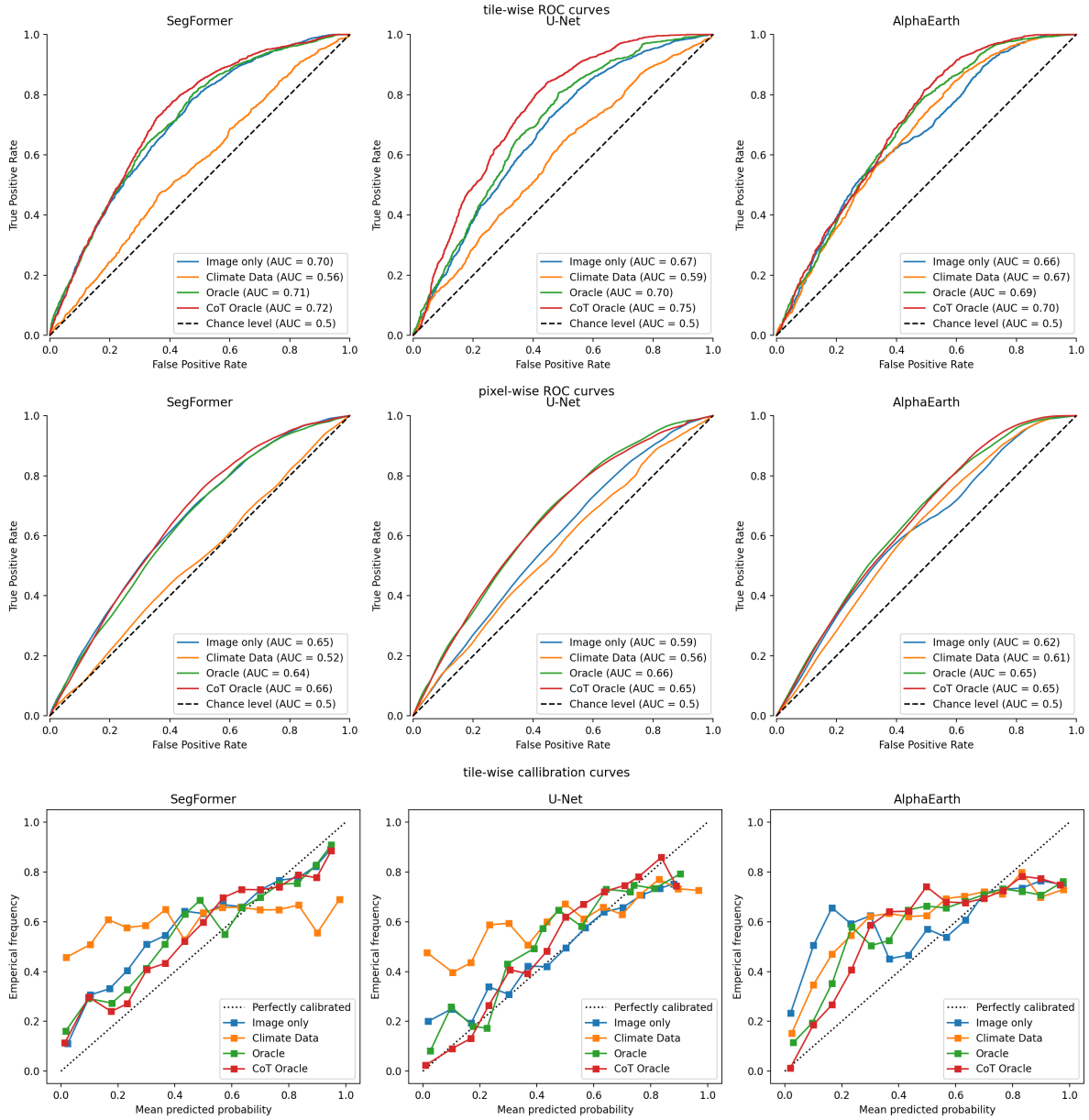
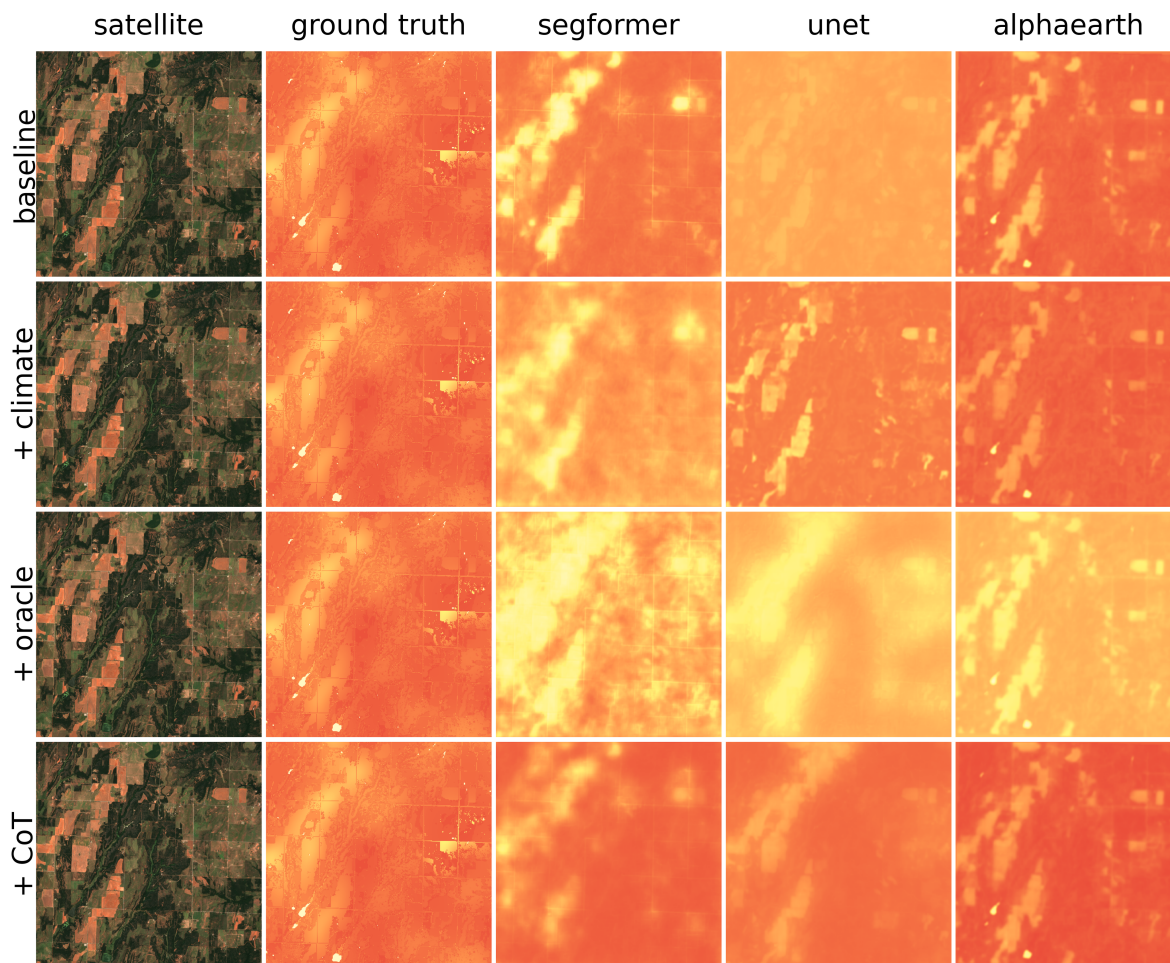


Figure 4. tile-wise ROC curves, pixel-wise ROC curves, tile-wise calibration curves



To determine the wildfire risk in the area based on the satellite image and climate data, we need to analyze the vegetation density, climate conditions, and wind patterns. Here's a detailed breakdown:

Vegetation Analysis:

- The satellite image shows a mix of agricultural fields and forested areas.
- The agricultural fields appear to be well-maintained, suggesting regular irrigation and management, which could reduce fire risk.
- The forested areas show signs of dryness, as indicated by the brownish color, which suggests that the vegetation might be more susceptible to fire.

Climate Data Analysis:

- **Average Temperature**: The average temperatures range from 2.59°C in January to 28.83°C in July. This indicates a significant variation in temperature throughout the year, with colder winters and warmer summers.
- **Average Wind Speed**: The wind speeds vary from 2.83 m/s in August to 3.97 m/s in April. Stronger winds can exacerbate fire spread.
- **Average Humidity**: The humidity levels are relatively low, especially in summer months like July and August, which can contribute to drier conditions and increased fire risk.
- **Average Precipitation**: The precipitation levels are generally low, with the highest being 3.94 mm per day in May. This suggests that the area may not receive much rain during the dry season, further increasing the risk of wildfires.

Wind Patterns:

- The wind directions indicate a variety of directions, but there is no consistent pattern that suggests a dominant direction that could align with the slopes and increase fire spread.
- The wind speeds are moderate, which is good for fire containment but could still be a concern if combined with dry vegetation and low humidity.

Risk Assessment:

- **Vegetation**: Mixed vegetation with both agricultural fields and forested areas, where the forested areas appear dry.
- **Climate**: Low humidity and moderate temperatures, which are conducive to dry conditions. However, the moderate wind speeds could help contain fires if they are not aligned with slopes.
- **Wind Patterns**: No consistent dominant wind direction, which reduces the risk of fire spread along a particular path.

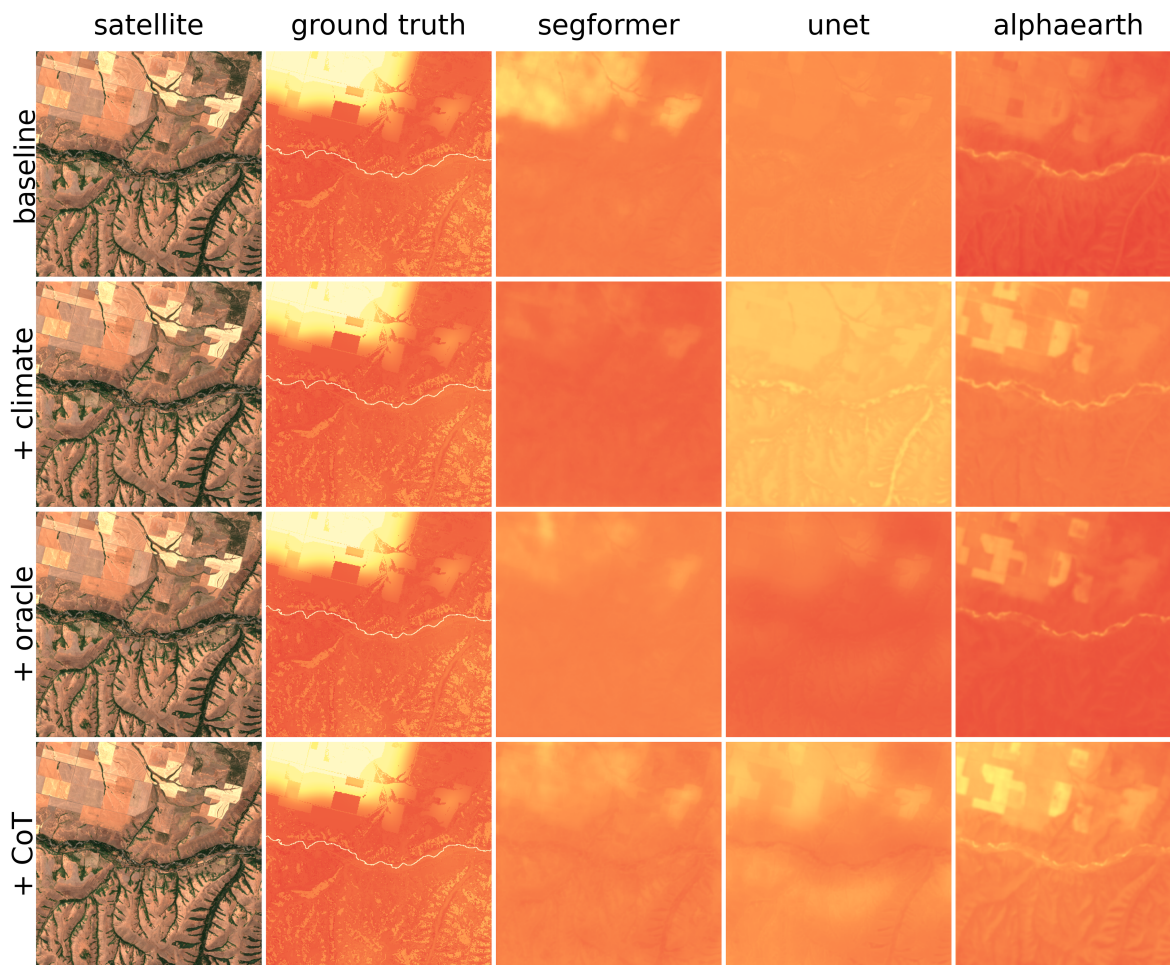
Final Classification:

Based on the analysis, the area has a moderate risk due to the presence of dry vegetation in some areas, low humidity, and moderate wind speeds. However, the lack of a consistent dominant wind direction mitigates the risk somewhat.

FINAL ANSWER:

7

Figure 5. 35.3996° N, -98.2942° W (Oklahoma).



To determine the wildfire risk in the area based on the satellite image and climate data, we need to analyze the following factors:

- Vegetation Density**: The satellite image shows a mix of agricultural fields and some green areas that could be vegetation. The presence of green areas suggests some vegetation, but the overall pattern indicates a significant amount of open land, which might not be as dense or continuous as needed for dense forest fires.
- Climate Data**:
 - Average Temperature**: The average temperatures range from -0.26°C to 21.47°C . While there are colder months, the warmer months (July and August) have temperatures above 20°C , which can contribute to drier conditions.
 - Average Humidity**: The humidity levels are relatively low, especially in July and August (45.78% and 47.68%), indicating dry conditions.
 - Average Wind Speed**: The wind speeds are moderate, ranging from 2.16 m/s to 3.2 m/s. However, the direction of the wind is crucial. The wind direction in July and August is 284.0° and 282.2° , which suggests a westerly direction. If this wind direction aligns with the slopes, it could exacerbate fire spread.
 - Average Precipitation**: The precipitation is very low, with most months having less than 2 mm per day, indicating dry conditions.
- Slope Alignment**: The satellite image shows a river running through the area, which could indicate a valley. The presence of a river suggests potential for slope alignment with the wind direction, which could enhance fire spread.

Given these factors, let's break down the risk level for different parts of the image:

- Top Left**: This area has a lot of agricultural fields, which are generally less flammable compared to dense vegetation. However, the dry climate and low humidity suggest a moderate risk.
- Top Middle**: This area also has agricultural fields, similar to the top left, suggesting a moderate risk.
- Top Right**: This area appears to have more mixed vegetation, including some green patches. The dry climate and low humidity still pose a risk, but the vegetation might slow the fire somewhat.
- Bottom Left**: This area has a mix of agricultural fields and some green areas. The dry climate and low humidity continue to pose a risk, but the vegetation might mitigate it slightly.
- Bottom Middle**: This area has a river, which could indicate a valley. The dry climate and low humidity suggest a high risk due to the potential for slope alignment with the wind.
- Bottom Right**: This area has a mix of agricultural fields and some green areas. The dry climate and low humidity continue to pose a risk, but the vegetation might mitigate it slightly.

Considering the overall pattern and the factors mentioned, the area appears to have a mix of moderate and high risk zones. The bottom middle section, which includes the river, seems to be the highest risk zone due to the potential for slope alignment with the wind.

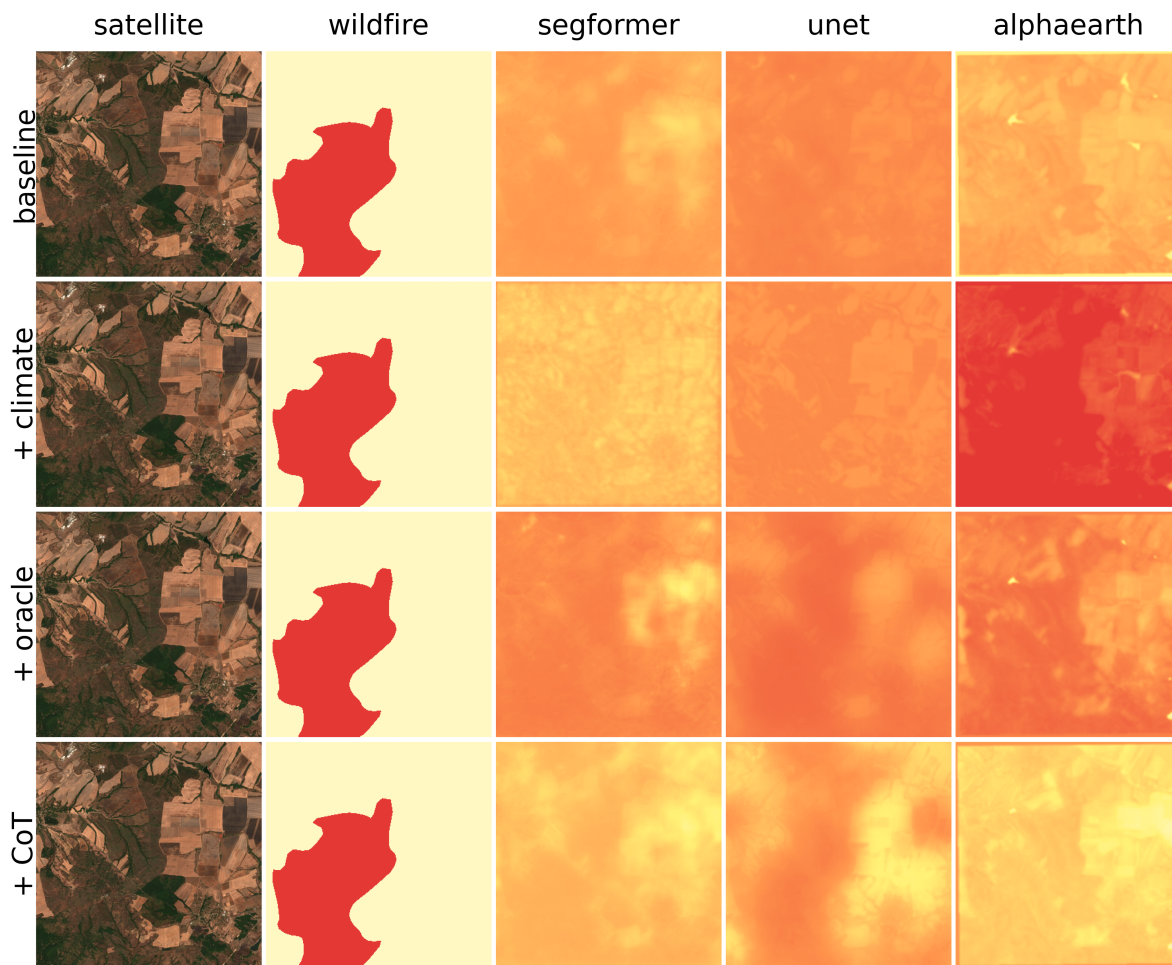
Final Answer:

The wildfire risk in the area ranges from moderate to high, with the bottom middle section being the highest risk zone. Based on this analysis, the overall risk level can be classified as follows:

FINAL ANSWER:

6

Figure 6. 45.6889°N , $-118.4442^{\circ}\text{W}$ (Oregon).



To determine the wildfire risk in the area based on the satellite image and climate data, we need to analyze the following factors:

1. **Vegetation Density**: The satellite image shows areas of dense vegetation (dark green) and less dense vegetation (lighter green). Dense vegetation can increase fire risk due to its higher fuel load.
2. **Climate Data**:
 - **Average Temperature**: Higher temperatures can increase the risk of wildfires.
 - **Average Wind Speed**: Strong winds can spread fires quickly.
 - **Average Humidity**: Lower humidity can exacerbate fire risk by reducing moisture in the vegetation.
 - **Average Precipitation**: Lower precipitation can reduce moisture levels in the soil and vegetation, increasing fire risk.

Analysis:

- **Vegetation Density**:
 - The dark green areas appear to be more dense vegetation, which could increase the fire risk.
 - The lighter green areas might indicate less dense vegetation, which could reduce fire risk.
- **Climate Data**:
 - **Temperature**: The average temperature in most months is relatively low, suggesting that the area is not experiencing extreme heat, which is beneficial for reducing fire risk.
 - **Wind Speed**: The wind speed is moderate but varies across the months. Strong winds can increase fire risk, especially if they align with slopes.
 - **Humidity**: The average humidity is generally high, indicating that the area is not experiencing very dry conditions, which would reduce fire risk.
 - **Precipitation**: The average precipitation is low, which could reduce moisture levels in the vegetation, increasing fire risk.

Risk Assessment:

- **Vegetation**: Mixed density, with some areas being dense and others less dense.
- **Temperature**: Generally low, which is good for fire risk.
- **Wind**: Moderate wind speeds, which can increase fire risk if aligned with slopes.
- **Humidity**: High humidity, which is good for fire risk.
- **Precipitation**: Low precipitation, which could increase fire risk.

Final Classification:

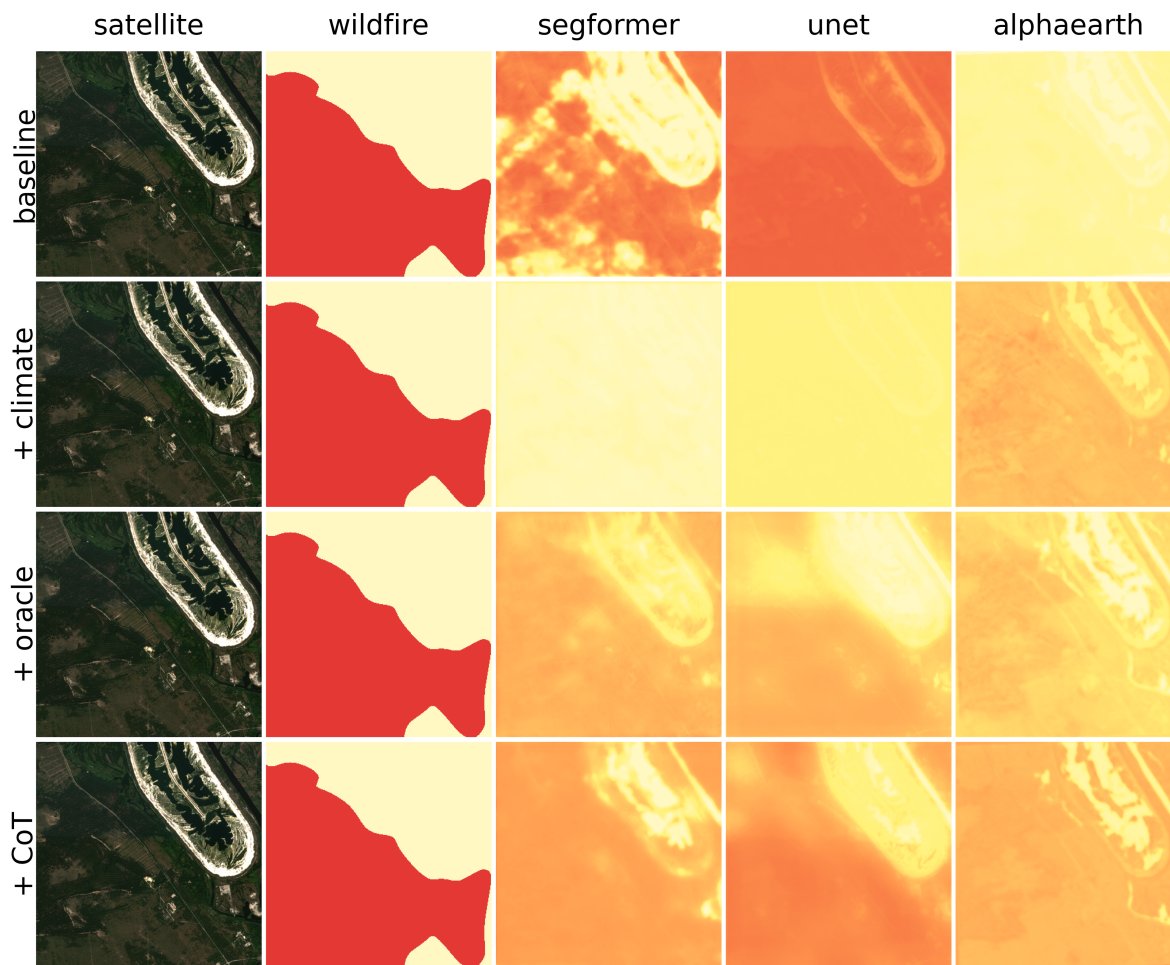
The area has mixed vegetation densities, with some dense areas that could increase fire risk. The climate data suggests that the area is not experiencing extreme heat or very dry conditions, which is beneficial. However, the low precipitation could increase fire risk. The moderate wind speeds could also increase fire risk if they align with slopes.

Given these factors, the overall risk appears to be moderate but with some areas of concern due to the dense vegetation and low precipitation.

FINAL ANSWER:

5

Figure 7. 42.1761° N, 26.161° W (Bulgaria). Fire event in 2020, pre-fire image from 2019.



To determine the wildfire risk in the area based on the satellite image and climate data, we need to analyze the vegetation type, land use, and climate conditions that influence fire risk. Here's a detailed breakdown:

Vegetation Analysis:

- **Top Right***: The area appears to be a coastal region with a sandy beach and some vegetation along the shoreline. This could be dune vegetation, which is generally less flammable but can still pose a risk if it becomes dry.
- **Bottom Left***: This area shows a mix of green and brown patches, indicating a combination of vegetation and possibly areas that have been burned or cleared. The brown patches suggest dry vegetation, which increases the risk of wildfires.
- **Center***: There are large patches of green, likely representing forested areas. These areas are more prone to wildfires due to dense vegetation.

Climate Data Analysis:

- **Average Temperature***: The average temperatures range from -6.62°C in January to 21.68°C in July. While there are colder months, the overall temperatures are relatively low, suggesting that the area might not experience extreme heat conducive to wildfires.
- **Average Wind Speed***: The wind speeds are moderate, ranging from 2.39 m/s in July to 3.21 m/s in March. Strong winds can exacerbate wildfires by spreading flames quickly, especially if they align with slopes.
- **Average Humidity***: The humidity levels are generally high, ranging from 59.65% in August to 95.14% in January. High humidity reduces the risk of wildfires as it makes the vegetation less flammable.
- **Average Precipitation***: The precipitation levels are low, ranging from 1.35 mm/day in April to 2.63 mm/day in July. Low precipitation indicates dry conditions, which can increase the risk of wildfires.

Risk Assessment:

- **Top Right***: Coastal areas with sandy beaches and some vegetation are less prone to wildfires due to the lack of dense vegetation and the presence of water nearby.
- **Bottom Left***: Areas with mixed vegetation and brown patches indicate dry vegetation, which increases the risk of wildfires.
- **Center***: Dense forested areas are highly susceptible to wildfires due to their dense vegetation.

General Classification:

Based on the analysis, the bottom left part of the image has the highest risk due to the presence of dry vegetation, while the top right part has a lower risk due to the sandy beach and less dense vegetation. The center part, though forested, does not appear to be at the highest risk because the climate data suggests moderate temperatures and high humidity, which reduce the risk of wildfires.

Final Answer:

The wildfire risk in the area ranges from low to moderate, with the bottom left part being the most at risk. A general classification for the wildfire risk in the area would be around 5 out of 9.

FINAL ANSWER:

5

Figure 8. 51.3168°N , 30.1658°W (Ukraine). Fire event in 2020, pre-fire image from 2019.

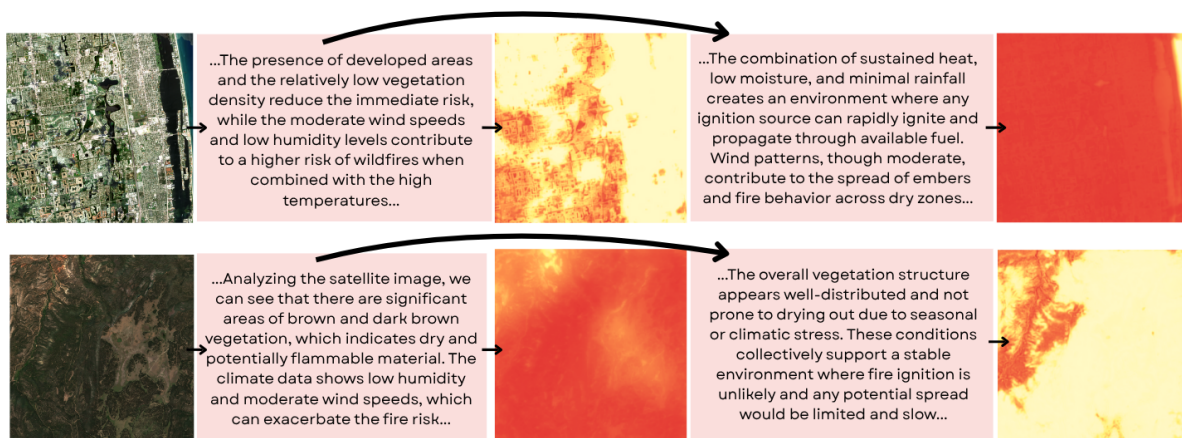


Figure 9. Visualization of U-Net FireScope’s adherence to its CoT and resulting high fidelity. After the CoT is artificially perturbed, the resulting raster changes accordingly.

References

- [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhao-hai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *preprint arXiv:2502.13923*, 2025. 2
- [2] G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. 1
- [3] Christopher F. Brown, Michal R. Kazmierski, Valerie J. Pasquarella, William J. Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, Noel Gorelick, Lihui Lydia Zhang, Sophia Alj, Emily Schechter, Sean Askay, Oliver Guinan, Rebecca Moore, Alexis Boukouvalas, and Pushmeet Kohli. AlphaEarth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *preprint arXiv:2507.22291*, 2025. 2
- [4] Jacob Cohen. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220, 1968. 2
- [5] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. 1
- [6] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. Soft actor-critic algorithms and applications. *preprint arXiv:1812.05905*, 2019. 3
- [7] P. Jaccard. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50, 1912. 2
- [8] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *preprint arXiv:1711.05101*, 2017. 4
- [9] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*, 2017. 4
- [10] M. P. Naeini, G. F. Cooper, and M. Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *AAAI Conference on Artificial Intelligence*, 2015. 1
- [11] OpenAI. Introducing GPT-5, 2025. Accessed: Nov. 12, 2025. 2
- [12] Marc-André Parisien and Max A. Moritz. Environmental controls on the distribution of wildfire at multiple spatial scales. *Ecological Monographs*, 79(1):127–154, 2009. 2
- [13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015. 2
- [14] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *preprint arXiv:1707.06347*, 2017. 3
- [15] Zhihang Shao, Ziyu Wang, Yuxin Zhang, Zihan Zheng, Yao Liu, Zihan Liu, Yibo Shang, Linyang Xu, Tianyang Zhang, Lingpeng Chen, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *preprint arXiv:2402.03300*, 2024. 3
- [16] USDA Forest Service and U.S. Geological Survey. Monitoring trends in burn severity (mtbs): Burned area boundaries dataset. <https://www.mtbs.gov>, 2024. USFS Geospatial Technology and Applications Center (GTAC), Salt Lake City, UT and USGS Earth Resources Observation and Science (EROS) Center, Sioux Falls, SD. 1
- [17] C. E. Van Wagner. Development and structure of the canadian forest fire weather index system. Technical Report Forestry Technical Report 35, Canadian Forestry Service, Petawawa National Forestry Institute, Chalk River, Ontario, 1987. 2
- [18] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 1
- [19] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2
- [20] Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhui Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2024*, pages 9556–9567, 2024. accessed: November 2025. 4, 7