**Descriptor-Guided Multi-Modal Wildfire Spread Forecasting (Research Proposal)**

**Background and Motivation**

**Challenges in Wildfire Forecasting:** Wildfire spread modeling is notoriously difficult due to the chaotic nature of fires and limitations in current sensing and prediction methods. Optical wildfire imagery is often **noisy and obscured by smoke**, which hides critical details of the fire front. Standard RGB cameras cannot see through heavy smoke plumes ([Seeing Through the Smokescreen: How Multi-Format Imagery Enables Effective Wildfire Response](https://resources.mckenzieintelligence.com/blog/seeing-through-the-smokescreen-how-multi-format-imagery-enables-effective-wildfire-response#:~:text=The%20naked%20eye%20cannot%20see,areas%20or%20predicting%20its%20movement)), making it nearly impossible to locate hotspots or predict fire movement using only visible light. Furthermore, deep learning models for wildfire detection/forecasting tend to operate as black boxes, offering **little interpretability** to human experts. Integrating explainability and transparency into wildfire AI models remains an open challenge ([Machine Learning and Deep Learning for Wildfire Spread Prediction: A Review](https://www.mdpi.com/2571-6255/7/12/482#:~:text=,improvement%20of%20datasets%20and%20metrics)). These limitations reduce the trust and usefulness of current systems in real-world firefighting, where understanding *why* a model predicts a certain spread is as important as accuracy.

([Seeing Through the Smokescreen: How Multi-Format Imagery Enables Effective Wildfire Response](https://resources.mckenzieintelligence.com/blog/seeing-through-the-smokescreen-how-multi-format-imagery-enables-effective-wildfire-response)) *A wildfire seen from above in (a) visible spectrum and (b) infrared. Thick smoke in (a) obscures the fire’s extent, whereas the IR composite in (b) renders smoke almost transparent and highlights active flames (bright areas) and burn scars (*[*Seeing Through the Smokescreen: How Multi-Format Imagery Enables Effective Wildfire Response*](https://resources.mckenzieintelligence.com/blog/seeing-through-the-smokescreen-how-multi-format-imagery-enables-effective-wildfire-response#:~:text=Shorter%20wavelengths%20present%20in%20visible,effectively%20rendering%20the%20smoke%20transparent)*). Such multi-modal data can reveal fire details that RGB alone misses.*

**Role of Infrared (IR) and Descriptors:** Infrared imaging offers a compelling solution to the smoke occlusion problem. IR wavelengths penetrate dense smoke, effectively rendering it transparent to the sensor ([Seeing Through the Smokescreen: How Multi-Format Imagery Enables Effective Wildfire Response](https://resources.mckenzieintelligence.com/blog/seeing-through-the-smokescreen-how-multi-format-imagery-enables-effective-wildfire-response#:~:text=Shorter%20wavelengths%20present%20in%20visible,effectively%20rendering%20the%20smoke%20transparent)). This reveals the underlying fire intensity and spread even when thick smoke is present, providing crucial information about flame fronts and hotspots that RGB images alone cannot capture. Indeed, studies have found that fusing thermal infrared data with visible imagery can significantly improve wildfire detection robustness and accuracy ([Sample images for the RGB-NIR and Corsican Fire Database datasets | Download Scientific Diagram](https://www.researchgate.net/figure/Sample-images-for-the-RGB-NIR-and-Corsican-Fire-Database-datasets_fig5_356175598#:~:text=Wildfire%20detection%20is%20of%20paramount,fusion%2C%20there%20is%20a%20gr)). Beyond raw sensor data, **high-level fire descriptors** (such as “smoke coverage = 80%” or “flame front moving north”) could further enhance forecasting. These descriptors serve as human-understandable features that describe the state of the fire. Current systems seldom leverage such intermediate representations, but **descriptor-guided learning** offers a path to better interpretability. In the machine learning literature, *concept bottleneck models* have shown that predicting human-interpretable concepts as an intermediate step can maintain predictive accuracy while enabling explanation in terms of those concepts ([[2007.04612] Concept Bottleneck Models](https://arxiv.org/abs/2007.04612#:~:text=concept%20bottleneck%20models%20by%20editing,on%20concepts%20at%20test%20time)). Applying this idea to wildfires, one can hypothesize that if a model explicitly recognizes attributes like smoke density or flame edge location, it may predict fire spread more reliably under noisy conditions (and explain its reasoning via those attributes).

**Advances in Vision-Language Models:** Another recent development motivating our approach is the rise of large vision-language models (VLMs) capable of richly describing images. Models like **GPT-4 Vision** (the multi-modal version of GPT-4) and **BLIP-2** can accept images and produce detailed text descriptions ([[2303.08774] GPT-4 Technical Report - arXiv](https://arxiv.org/abs/2303.08774#:~:text=We%20report%20the%20development%20of,inputs%20and%20produce%20text%20outputs)) ([[2301.12597] BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models](https://arxiv.org/abs/2301.12597#:~:text=image%20encoder,can%20follow%20natural%20language%20instructions)). For example, a VLM can look at a wildfire image and output descriptions like “billowing dark smoke covering the left half of the fire, with bright flames visible at the base, moving uphill.” Such automatically generated descriptors could be invaluable for guiding a forecasting model. They provide a form of *interpretable signal* extracted from the imagery. By leveraging VLMs, we can generate candidate descriptors at scale, rather than hand-crafting rules. Additionally, these descriptors offer a way to integrate expert knowledge or **human feedback** into the loop – experts can verify or correct them, making the model’s inner reasoning more transparent. Overall, the availability of multi-modal data (RGB + IR) and powerful image description models presents a timely opportunity: we can design an AI pipeline that fuses traditional visual inputs with **descriptor-based features** to overcome data noise and improve both the accuracy and interpretability of long-term wildfire spread forecasts.

**Research Objectives**

Our research will pursue the following key objectives:

* **O1: Multi-Modal Forecasting Pipeline** – Develop a novel AI pipeline that **fuses RGB imagery, IR imagery, and fire descriptors** to predict wildfire spread. The pipeline should ingest aligned RGB and IR frames and output forecasts (e.g. next fire location or growth) while internally utilizing descriptor features (like smoke density, flame presence, spread direction).
* **O2: Descriptor vs. Image Modalities** – Rigorously evaluate the contribution of the intermediate descriptors. We will compare model variants using: (a) images only, (b) descriptors only, and (c) combined images + descriptors. This will test whether **descriptors alone (textual features)** can outperform purely image-based models on certain tasks, and whether adding descriptors to image models improves performance beyond images alone.
* **O3: Vision-Language Descriptor Generation** – Assess the use of large vision-language models to automatically generate wildfire descriptors from imagery. We will test models like GPT-4V, BLIP-2, and Google’s Gemini on fire images to produce descriptive captions. The **accuracy and domain relevance** of these AI-generated descriptors will be evaluated against ground truth and expert judgment.
* **O4: Temporal Modeling for Spread Forecasting** – Design and implement temporal deep learning models for wildfire spread prediction that operate on sequences of multi-modal data. This includes constructing baseline architectures (e.g. a CNN-LSTM that ingests image sequences) and advanced architectures (e.g. a Transformer-based sequence model) to forecast fire progression over time. We will explore different **fusion strategies** for combining RGB, IR, and descriptor inputs over time (early fusion vs. late fusion, discussed below).
* **O5: Interpretability and Explainability** – Throughout, evaluate if the introduction of descriptors improves the interpretability of the model’s predictions. We aim to show that our system can provide human-understandable reasoning (e.g. “prediction: fire will spread north, because descriptor = strong winds from south and visible flame at northern edge”). Even if overall accuracy is on par with black-box models, this improved explainability would be a valuable outcome.

By achieving these objectives, the project will demonstrate the feasibility and benefits of **descriptor-guided, multimodal learning** for a critical real-world forecasting problem.

**Methodology**

Our methodology is organized into several components, from data preparation to model development and evaluation:

* **Dataset and Ground Truth Annotation:** We will leverage the **Corsican Fire Database (CFDB)** as the primary dataset ([CORSICAN FIRE DATABASE | Projet Feux | Università di Corsica Pasquale Paoli | Université de Corse Pasquale Paoli](https://feuxdeforet.universita.corsica/article.php?id_art=2133&id_rub=572&id_menu=0&id_cat=0&id_site=33&lang=en#:~:text=This%20database%2C%20named%20the%20Corsican,range%20and%20near%20infrared%20range)). This public dataset contains wildfire imagery captured in the **visible (RGB) spectrum and the near-infrared (thermal)** spectrum, with many images available as paired RGB-IR observations of the same scene ([CORSICAN FIRE DATABASE | Projet Feux | Università di Corsica Pasquale Paoli | Université de Corse Pasquale Paoli](https://feuxdeforet.universita.corsica/article.php?id_art=2133&id_rub=572&id_menu=0&id_cat=0&id_site=33&lang=en#:~:text=This%20database%2C%20named%20the%20Corsican,range%20and%20near%20infrared%20range)). Notably, CFDB also provides a limited number of **temporal sequences** (video frame sets) of fires captured simultaneously in RGB and IR ([CORSICAN FIRE DATABASE | Projet Feux | Università di Corsica Pasquale Paoli | Université de Corse Pasquale Paoli](https://feuxdeforet.universita.corsica/article.php?id_art=2133&id_rub=572&id_menu=0&id_cat=0&id_site=33&lang=en#:~:text=This%20database%2C%20named%20the%20Corsican,range%20and%20near%20infrared%20range)) ([Computer vision for wildfire research: An evolving image dataset for processing and analysis | Request PDF](https://www.researchgate.net/publication/318284024_Computer_vision_for_wildfire_research_An_evolving_image_dataset_for_processing_and_analysis#:~:text=issue%20by%20presenting%20a%20publicly,released%20in%20this%20research%20field)), which is ideal for our forecasting tasks. Each image in CFDB is accompanied by meta-data such as the manually outlined fire perimeter, dominant flame color, percentage of the fire region obscured by smoke, etc. ([CORSICAN FIRE DATABASE | Projet Feux | Università di Corsica Pasquale Paoli | Université de Corse Pasquale Paoli](https://feuxdeforet.universita.corsica/article.php?id_art=2133&id_rub=572&id_menu=0&id_cat=0&id_site=33&lang=en#:~:text=distance%20to%20the%20fire%20and,the%20brightness%20of%20the%20environment)). These annotations will inform which descriptors are relevant. We will **manually label a subset** of ~150 images with a concise set of **wildfire-specific descriptors** to use as ground truth for descriptor prediction. Example descriptors include: “Smoke coverage (%)”, “Flame front visible (y/n)”, “Direction of main spread (N, S, E, W)”, “Multiple fronts present (y/n)”, “Embers spotting beyond main fire (y/n)”, etc. This ontology is inspired by the CFDB annotations (e.g. smoke coverage) and wildfire expert knowledge. The manual labels (by domain experts if possible) will serve to both train a descriptor prediction model and to evaluate the accuracy of AI-generated descriptors.
* **Automated Descriptor Generation (Vision-Language Models):** We will apply state-of-the-art **vision-language models** to generate descriptive captions or labels from the images. In particular, we plan to test **GPT-4 Vision** (via the GPT-4 multimodal API) and **BLIP-2** ([[2301.12597] BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models](https://arxiv.org/abs/2301.12597#:~:text=image%20encoder,can%20follow%20natural%20language%20instructions)), and possibly **Google Gemini** (a recent multimodal LLM ([[2312.11805] Gemini: A Family of Highly Capable Multimodal Models](https://arxiv.org/abs/2312.11805#:~:text=,device))), on our fire images. Each model will be prompted to output a structured description focusing on wildfire characteristics (we will craft prompts such as *“Describe the wildfire in this image, including smoke, flames, and direction of spread.”*). The result will be one or more sets of AI-generated descriptors for each image. We will then **compare the quality of these descriptors** against our ground truth labels using both automated metrics and human judgment. Automated metrics will include **BLEU score** (to measure n-gram overlap between generated text and reference descriptors) and **CLIPScore**, a learned metric that correlates well with human caption quality ([CLIPScore: A Reference-free Evaluation Metric for Image Captioning - ACL Anthology](https://aclanthology.org/2021.emnlp-main.595/" \l ":~:text=image%2Bcaption%20pairs%20from%20the%20web%2C,Beyond%20literal%20description%20tasks)). CLIPScore works by embedding the image and the generated description in a joint vision-language space and measuring their compatibility ([CLIPScore: A Reference-free Evaluation Metric for Image Captioning - ACL Anthology](https://aclanthology.org/2021.emnlp-main.595/" \l ":~:text=image%2Bcaption%20pairs%20from%20the%20web%2C,Beyond%20literal%20description%20tasks)) – a higher score means the description is more on-point for the image. Additionally, we will use a **CLIP-based similarity** to check if key terms (like “smoke” or “flame”) are appropriately mentioned. Importantly, we will have wildfire experts perform a **human evaluation**: they will rate whether the descriptors are factually correct and relevant to wildfire behavior. This evaluation will reveal which model (GPT-4V vs BLIP-2 vs others) produces the most **accurate and domain-relevant** descriptions of wildfires. Those descriptions that are sufficiently accurate will be used in subsequent model training; we may also consider combining multiple VLM outputs or using an ensemble approach for robustness.
* **Baseline Models (Image-Only vs. Descriptor-Only):** Before fusing everything, we will establish baseline performances for predicting wildfire attributes and spread using single modalities:
  + *Image-only baselines:* We will train deep CNN models on the RGB images (and separately on IR images) to perform tasks like fire classification and next-frame prediction. For image classification (e.g. predicting if the fire will spread left vs right), a high-performing CNN such as **EfficientNet** will be used. EfficientNet is a family of CNNs that scale in depth/width while maintaining high accuracy efficiently ([EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks](https://www.researchgate.net/publication/333444574_EfficientNet_Rethinking_Model_Scaling_for_Convolutional_Neural_Networks" \l ":~:text=,)) – these models have achieved top accuracy on image recognition benchmarks with relatively fewer parameters. We will fine-tune an EfficientNet on our dataset for tasks like detecting presence of smoke or predicting immediate spread direction. We will also experiment with a **Vision Transformer (ViT)** model ([[2010.11929] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale](https://arxiv.org/abs/2010.11929#:~:text=reliance%20on%20CNNs%20is%20not,fewer%20computational%20resources%20to%20train)), given their success in image recognition by using transformer attention on image patches. ViTs have shown excellent results on image classification, rivaling CNNs ([[2010.11929] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale](https://arxiv.org/abs/2010.11929#:~:text=reliance%20on%20CNNs%20is%20not,fewer%20computational%20resources%20to%20train)), which makes them a strong baseline for capturing spatial features of fire images. For sequence modeling (predicting future frames), a simple image-only baseline will be a **CNN-LSTM**: e.g., an EfficientNet (for spatial features) whose output features at each time-step are fed into an LSTM to capture temporal patterns. This is a common approach for video sequence prediction tasks.
  + *Descriptor-only baselines:* Here we treat the problem as a pure NLP/text classification or sequence problem using only the high-level descriptors (no raw images). We will train a model on **human-labeled descriptors** (and possibly synthetic descriptors from VLMs, if accurate) to predict outcomes like next-step fire spread. For example, given a series of textual descriptors over time (like “time1: heavy smoke, winds N; time2: flames visible moving north”), can we predict the next descriptor or next fire state? We will employ **BERT**, a transformer-based language model ([BERT (language model) - Wikipedia](https://en.wikipedia.org/wiki/BERT_(language_model)#:~:text=BERT%20%28language%20model%29%20,Wei%20Chang)), to encode the descriptors. BERT is well-suited since it provides deep bidirectional understanding of text ([BERT (language model) - Wikipedia](https://en.wikipedia.org/wiki/BERT_(language_model)#:~:text=BERT%20%28language%20model%29%20,Wei%20Chang)). A simple BERT-based classifier will be fine-tuned to predict classes such as “fire will grow vs. shrink in next frame” or “spread direction = north/south/etc.” from the text input. This baseline will tell us how informative the descriptors alone are. If the descriptor-only model performs reasonably, it validates our hypothesis that these semantic cues contain predictive signal. If it fails, it suggests descriptors alone are insufficient, underscoring the need to combine them with visual data.
* **Multi-Modal Fusion Models:** The core of our methodology is designing models that **integrate RGB, IR, and descriptors** for improved forecasting. We will investigate two primary architectures for temporal forecasting:
  1. **CNN-LSTM with Fusion:** In this approach, we extend the CNN-LSTM baseline to handle multi-modal input. The RGB and IR image at each time-step will be processed by two separate CNN backbones (or a dual-branch network) to extract features from each modality. We will then fuse these features along with descriptor information before feeding into the LSTM (which models the temporal evolution). We will experiment with **early fusion vs. late fusion**:
     + *Early Fusion:* Directly concatenate the raw input channels or early features. For example, we can concatenate the RGB image and IR image to form a 6-channel input (3 RGB channels + 3 IR channels) and feed this into a single CNN as if it were a 6-channel image ([Two approaches for feature fusion. | Download Scientific Diagram](https://www.researchgate.net/figure/Two-approaches-for-feature-fusion_fig3_365480586#:~:text=,)). This was found to be a simple yet effective way to fuse paired RGB-IR data in prior fire detection research ([Two approaches for feature fusion. | Download Scientific Diagram](https://www.researchgate.net/figure/Two-approaches-for-feature-fusion_fig3_365480586#:~:text=,)). We can analogously concatenate descriptor features (which could be a vector encoding from BERT) with the image feature vector at each time step.
     + *Late Fusion:* Keep modalities separate through most of the network and combine at a higher decision level. For instance, we could have one CNN-LSTM processing RGB, another processing IR, and a BERT (or another LSTM) processing the descriptor sequence; then their outputs are concatenated or merged by a fully-connected layer just before the final prediction. This allows each modality-specific branch to learn its own representation, merging only when making the forecasting decision.
     + *Middle Fusion:* (We will also consider an intermediate fusion, e.g. merging features after a few CNN layers, as suggested in some RGB-T fusion studies ([Big data and artificial intelligence in cancer research: Trends in ...](https://www.cell.com/trends/cancer/fulltext/S2405-8033(23)00217-0?rss=yes#:~:text=,L%2C%20network%27s%20layer%3B%20SM)), but primary focus will be early vs. late for simplicity.)

Regardless of fusion timing, we will ensure the model architectures are comparable in capacity. The training will involve feeding in sequences (e.g. 5-10 consecutive frames) and training the network to predict a future outcome, such as the fire segmentation map at the next time step or the movement direction of the fire front.

* 1. **Transformer-based Sequence Model:** Given the growing evidence that **Transformers can handle long-range dependencies in environmental sequences effectively (**[**Machine Learning and Deep Learning for Wildfire Spread Prediction: A Review**](https://www.mdpi.com/2571-6255/7/12/482#:~:text=)**)**, we will design a transformer model for our task. One concept is to use a **Vision Transformer encoder** for each modality frame and then a **Temporal Transformer** to model the sequence. Concretely, at each time step we obtain: an embedding from RGB (e.g. via a ViT), an embedding from IR, and an embedding from the descriptor text (via BERT). We then concatenate these modality embeddings into a single combined token per time step, and feed the sequence of combined tokens into a **Temporal Transformer** (similar to those used in video action recognition). The transformer’s self-attention will implicitly learn to weigh information from different modalities and different time steps. We will investigate positional encoding schemes to handle the time dimension. If data is limited, we may opt for a simpler Transformer that operates on the time axis after early fusing the modalities (to reduce complexity). Recent work by Qayyum et al. used a transformer for wildfire spread prediction with success ([Machine Learning and Deep Learning for Wildfire Spread Prediction: A Review](https://www.mdpi.com/2571-6255/7/12/482#:~:text=Qayyum%20et%20al.%20,the%20model%20outputs%20across%20different)), and our model will draw on similar ideas, with the addition of multi-modal inputs. We will also integrate an **attention-based interpretability mechanism**: for example, using attention weights to see which descriptors or which time steps the model focuses on when predicting the outcome (this can provide insight, e.g. the model might attend strongly to “windy” descriptor at time t when predicting spread at time t+1, which aligns with human intuition).
* **Training Strategy:** All models will be trained using a combination of classification and regression objectives. For static predictions (e.g. “will the fire spread to area X by next frame?” – a binary classification), we will use a cross-entropy loss. For spatial predictions (e.g. predicting the segmented fire mask for the next time step), we will use a segmentation loss such as Dice loss or IoU-based loss, which directly optimizes spatial overlap. The descriptor generation models (LLMs) may be used in a **two-stage training**: first stage, train/fine-tune the descriptor generator on the labeled subset to better fit wildfire terminology; second stage, use the frozen generator to supply descriptors for the full dataset to train the fusion model. We will be careful to avoid data leakage between the descriptor evaluation and model training (i.e., use separate sets for evaluating descriptor quality vs. using them as features in the forecasting model). Because the dataset is relatively small (hundreds of images, a handful of sequences), we will rely on pre-trained models (e.g. ImageNet pre-trained EfficientNet/ViT and BERT) and use **transfer learning** rather than training from scratch. We will perform data augmentation (random flips, rotation, mild color jitter for RGB images) to increase effective data and also possibly generate *simulated* IR images for augmentation using a learned visible-to-IR translation model (as noted in Future Work). Model selection will be done via cross-validation on the available sequences (e.g. leave-one-fire-out validation, where we train on sequences from some fires and test on a sequence from a different fire event, to judge generalization).
* **Fusion Strategy Experiments:** We will conduct ablation experiments to determine which fusion strategy yields the best results. For example, we will train one model with early fusion of RGB+IR at the image level, and another with late fusion combining separate RGB and IR predictions. Similarly, we’ll test fusion of descriptors at different points: feeding descriptors in as an additional “channel” vs. combining only at the decision layer. These experiments will be evaluated on a validation set to see which approach optimally balances performance and simplicity. Prior research in RGB-thermal fusion often found that mid-level feature fusion can outperform naive early concatenation ([Big data and artificial intelligence in cancer research: Trends in ...](https://www.cell.com/trends/cancer/fulltext/S2405-8033(23)00217-0?rss=yes#:~:text=,L%2C%20network%27s%20layer%3B%20SM)), but the optimal method can be task-dependent. We will quantitatively compare these on our forecasting metrics. The result will inform the design of our final proposed pipeline.

**Evaluation Plan**

We will evaluate the pipeline on both **static prediction tasks** (predicting attributes or immediate outcomes) and **temporal forecasting tasks** (predicting fire spread over time), using a suite of metrics for accuracy and interpretability:

* **Descriptor Generation Evaluation:** As mentioned, the quality of AI-generated descriptors will be evaluated by BLEU score (for textual overlap) and by CLIPScore (for alignment with image content) ([CLIPScore: A Reference-free Evaluation Metric for Image Captioning - ACL Anthology](https://aclanthology.org/2021.emnlp-main.595/" \l ":~:text=image%2Bcaption%20pairs%20from%20the%20web%2C,Beyond%20literal%20description%20tasks)). Additionally, we will report **precision/recall** of key terms in the generated descriptions (for example, if ground truth says “smoke present” and model says “no smoke”, that’s a miss). A small expert user study will provide qualitative scores on descriptiveness and correctness. These results will be reported to justify the choice of descriptor generator model in the final pipeline.
* **Static Classification Metrics:** For tasks like detecting if an image has heavy smoke, or whether the fire will move in a certain direction (treated as classification problems), we will use standard metrics: **Accuracy, Precision, Recall, F1-score**. Accuracy alone can be misleading for imbalanced outcomes, so we will emphasize precision and recall (e.g. high recall means we rarely miss a spreading-fire warning). These metrics are common in wildfire prediction literature ([Machine Learning and Deep Learning for Wildfire Spread Prediction: A Review](https://www.mdpi.com/2571-6255/7/12/482#:~:text=quantify%20the%20error%20in%20the,prediction%20of%20burned%20areas)). For example, if we classify each frame as “will the fire grow in next frame or not”, we will compute the confusion matrix and derive precision/recall of growth predictions. We will compare these metrics between the image-only model vs. descriptor-only vs. fused model to see which is best. An improvement in F1 when using descriptors would support our hypothesis that descriptors add useful information.
* **Spatial Prediction Metrics:** When evaluating the **forecasted fire spread area** (e.g. as a segmentation mask or burned area prediction), we will use spatial overlap measures like **Intersection over Union (IoU)**. IoU measures how well the predicted fire region matches the ground truth region by the size of their overlap divided by the size of their union ([Machine Learning and Deep Learning for Wildfire Spread Prediction: A Review](https://www.mdpi.com/2571-6255/7/12/482#:~:text=,models%20used%20in%20wildfire%20spread)). This is a standard metric for comparing predicted vs. actual burned areas in image-based wildfire models ([Machine Learning and Deep Learning for Wildfire Spread Prediction: A Review](https://www.mdpi.com/2571-6255/7/12/482#:~:text=,models%20used%20in%20wildfire%20spread)). We will also use the **Hausdorff Distance** to evaluate the accuracy of predicted fire front contours. Hausdorff Distance measures the maximum distance between the predicted fire boundary and the true boundary ([Modelling forest fire dynamics using conditional variational autoencoders | Information Systems Frontiers](https://link.springer.com/article/10.1007/s10796-024-10507-9#:~:text=Hausdorff%20Distance,In%20simpler) ); a smaller Hausdorff distance means the shape and position of the fire front was predicted more precisely. This metric is useful for judging how well the model predicts the *shape* of the fire, not just its area. Prior work on data-driven fire spread simulation has employed Hausdorff Distance for comparing fire perimeters ([Modelling forest fire dynamics using conditional variational autoencoders | Information Systems Frontiers](https://link.springer.com/article/10.1007/s10796-024-10507-9#:~:text=Table%204%20Similarity%20Evaluation,based%20sampling) ), so it is suitable here. For each method (RGB-only vs. IR-only vs. fused), we will compute the IoU and Hausdorff of the predicted fire mask at various forecast horizons (e.g. 1 frame ahead, 3 frames ahead, etc.). We expect the combined RGB+IR+Descriptor model to achieve higher IoU (meaning it better anticipates where the fire will be) than models lacking one of these inputs.
* **Temporal Sequence Metrics:** To evaluate longer-term forecasts (e.g. predicting fire spread 5 steps into the future), we will look at sequence prediction metrics. One is the **mean IoU over the sequence** – essentially averaging the overlap of predicted vs actual fire area for each time step in the forecast horizon ([Modelling forest fire dynamics using conditional variational autoencoders | Information Systems Frontiers](https://link.springer.com/article/10.1007/s10796-024-10507-9#:~:text=Table%206%20End,within%20the%20video%20timeline) ). We will also measure the **time to error** – how many time-steps into the future until the prediction fails catastrophically (IoU falls below a threshold, or the predicted fire location deviates by more than a certain distance). This gives a sense of how far out our model can predict with reliable accuracy. If possible, we’ll employ the **Hausdorff distance at the final forecast frame** as an indicator of worst-case boundary error over the prediction horizon ([Modelling forest fire dynamics using conditional variational autoencoders | Information Systems Frontiers](https://link.springer.com/article/10.1007/s10796-024-10507-9#:~:text=match%20at%20L1045%20Table%206,within%20the%20video%20timeline) ). These temporal metrics will help compare the CNN-LSTM vs. Transformer approaches: Transformers might maintain accuracy further into the future due to better sequence modeling, which would reflect in higher mean IoU at longer horizons.
* **Qualitative Interpretability Assessment:** Since interpretability is a major goal, we will perform a qualitative evaluation of the system’s outputs. This will involve examining example predictions where the model uses descriptors and seeing if the reasoning is understandable. For instance, we will look at cases where the model predicts “fire will intensify” and check if the descriptors that the model relied on (explicitly or implicitly) make sense (perhaps the model noted “increasing smoke and embers downwind” which aligns with intensification). We will generate **saliency maps** for the image inputs (to see which parts of RGB/IR frames the model focused on) and attention weight visualizations for descriptor tokens in the Transformer model. If our approach is successful, we expect to observe that, say, the model’s attention is drawn to words like “strong wind” or image regions of thick smoke when predicting a rapid spread – an interpretable correlation. We will present such examples in the results to illustrate how the descriptors contribute to model decisions. Additionally, we plan to interview a domain expert with our model outputs: showing them the descriptor-guided predictions vs. a black-box model’s predictions, and gathering feedback on which they trust more and why.
* **Simulation and Case Study:** As a final evaluation, we will test the trained pipeline on a **simulated fire progression** scenario using the Corsican dataset. For example, we might take an image of an early-stage fire and then simulate a progression by feeding the model repeatedly (using its own output iteratively). We can utilize the small number of video sequences in CFDB as ground truth for this – e.g., take the first frame of a sequence, have the model predict frame 2, then use predicted frame 2 to predict frame 3, and so on, and compare against the real sequence frames. This will really stress-test the model’s long-term predictive ability. We will measure how well it maintains accuracy over multiple iterative predictions (this often causes error accumulation, a known challenge in forecasting). We might quantify the **multi-step prediction drift** (difference between actual and predicted fire perimeter after N steps). Successful performance here would demonstrate the model’s practical utility in forecasting a fire’s growth trajectory from an initial snapshot. All evaluations will be performed separately on at least two conditions: daytime vs. nighttime imagery (as IR is especially crucial at night, we expect our IR-enhanced model to shine there), and possibly on different terrain types if data allows (forest vs. grassland fires).

The evaluation results will be compiled to compare: (i) RGB-only vs IR-only vs descriptors-only vs combined, (ii) early vs late fusion, (iii) CNN-LSTM vs Transformer, and (iv) human vs AI-generated descriptors. Statistical significance tests (where applicable) will be used to confirm improvements (e.g., McNemar’s test for classification accuracy differences). We anticipate that the **RGB+IR+Descriptor fused model will outperform others on most accuracy metrics**, and we will carefully analyze any cases where it does not (e.g., if descriptors are noisy, the image-only model might win in some instances – which would highlight the importance of descriptor quality).

**Future Work and Extensions**

Beyond the scope of the initial proposal, there are several promising directions to extend this research:

* **Incorporating Environmental and Geospatial Data:** In addition to image-based descriptors, future versions of the model could ingest **contextual metadata** like wind speed & direction, humidity, temperature, and terrain maps. Such factors critically influence fire spread (strong winds can drive fires quickly, certain fuel types burn faster, etc.), and integrating them could improve forecasting accuracy. Many current wildfire spread datasets lack this information or have it at low resolution ([Machine Learning and Deep Learning for Wildfire Spread Prediction: A Review](https://www.mdpi.com/2571-6255/7/12/482#:~:text=Datasets%20used%20in%20wildfire%20spread,discrepancies%20among%20sources%2C%20which%20complicate)). We propose as future work to pair the image data with coarsely sampled weather data (from nearby weather stations or reanalysis) and with topographical features (slope, aspect). These can be fed as additional descriptors or as extra input channels to the model. We expect that adding this could turn our model into a more holistic system that understands not just what it “sees” but also the underlying drivers of fire behavior (for example, the model could learn that a descriptor “wind blowing east at 20 km/h” combined with seeing flames means a high chance of eastern spread).
* **Real-Time Deployment and Edge Optimization:** Ultimately, the goal is to deploy wildfire forecasting in real operational settings (e.g., mounted cameras or drone feeds monitoring forests). Future work will focus on **optimizing the pipeline for real-time inference**. This involves model compression or using lightweight architectures so that the model can run on edge devices or with low latency. The need for real-time, low-computation models in dynamic wildfire events has been emphasized in previous studies ([Machine Learning and Deep Learning for Wildfire Spread Prediction: A Review](https://www.mdpi.com/2571-6255/7/12/482#:~:text=there%20are%20some%20challenges%20related,addition%2C%20ensuring%20that%20the%20model%E2%80%99s)). We might explore quantization or knowledge distillation to reduce model size. Another aspect is an **incremental update scheme**: instead of re-processing a long sequence for each new frame, maintain a state (like the LSTM’s hidden state or the Transformer’s encoded memory) that can be updated as new frames arrive. This would allow continuously running forecasts with minimal overhead per new frame. In field deployment, the system would take incoming camera frames, generate descriptors (possibly on-device via a small model like BLIP-2 with a smaller language head), and then feed everything into the spread predictor to give firefighters an updated forecast every few minutes. Future research can demonstrate a prototype of this, evaluating it in streaming mode.
* **Expansion to Satellite and Aerial Imagery:** While CFDB is ground-based, an exciting extension is to apply the approach to **satellite or aerial wildfire imagery**, such as data from the VIIRS or MODIS satellites (which provide IR detection of fires) or from drone overflights. Satellite images cover larger areas and could allow forecasting the spread of very large fires, though at lower resolution. The descriptor approach could be even more useful there, since human analysts often annotate satellite fire data with terms like “fire intensity” or “plume size” qualitatively. We would need to handle differences in perspective and scale, but the multi-modal pipeline concept should transfer. A possible future study is to use **NASA Fire Information for Resource Management System (FIRMS)** data (which gives satellite fire detections and could be treated as “IR hotspots”) fused with optical imagery from Landsat or Sentinel satellites for the same fires. This would broaden the impact of our work from local cameras to wide-area monitoring.
* **Synthetic Data Generation (RGB→IR image translation):** One practical hurdle is that IR cameras are not always available for all fires – but RGB cameras are common. To mitigate this, future work can develop a model to **generate synthetic IR images from RGB** input. Using generative adversarial networks (GANs) or emerging diffusion models, we could translate a visible image to its thermal-infrared equivalent. Preliminary research (e.g. InfraGAN) has shown it’s feasible to produce an IR-esque image from an RGB image by learning the mapping on training pairs ([InfraGAN: A GAN architecture to transfer visible images to infrared ...](https://www.sciencedirect.com/science/article/abs/pii/S0167865522000332" \l ":~:text=InfraGAN%3A%20A%20GAN%20architecture%20to,by%20training%20our%20deep%20network)). If successful, this would allow our pipeline to be used even when only RGB is present: the GAN would produce a pseudo-IR image, which our forecasting model (trained on real IR) could then use. This effectively augments the data and addresses missing modality issues. Our future plan is to train such a translation model on the CFDB’s 100 RGB-IR image pairs (possibly expanding with other datasets like FLAME2 that also have paired data) and integrate it as a preprocessing step. Even if the synthetic IR is not perfect, it could provide additional signal (e.g., highlighting likely hotspots through smoke) to the spread model. We will need to carefully evaluate the impact of using generated IR – ensuring it doesn’t introduce bias or artifacts. If it proves beneficial, this approach can greatly extend the applicability of multimodal forecasting, since any standard camera could be “upgraded” with a virtual IR channel.
* **Generalization to Other Fire Regimes and Climate Conditions:** Wildfires vary (forest fires vs grass fires, etc.), and our initial work focuses on the Corsican dataset conditions (Mediterranean ecosystem fires). A valuable extension is to test and adapt the pipeline to other wildfire datasets and even to synthetic wildfire simulations. Future research could incorporate data from North American fires, Australian bushfires, etc., to ensure the system generalizes. We might need to incorporate a broader range of descriptors (for example, fires in peatlands have thick smoke of different character, etc.). The architecture could also be extended to forecast not just spread area but other quantities like smoke dispersion or flame height if those descriptors are added – making it a more comprehensive wildfire behavior model.

In summary, the future work will push the model towards real-world deployment: adding more data sources (environmental sensors), making it real-time and widely usable (through sensor fusion and synthetic data), and broadening its coverage to various wildfire scenarios. Each extension would move us closer to a field-ready AI assistant for wildfire management.

**Expected Contributions**

If successful, this research will yield several important contributions to the fields of wildfire modeling and multi-modal AI:

* **Novel Descriptor-Guided Pipeline:** We will introduce a first-of-its-kind pipeline for **multimodal wildfire spread forecasting** that tightly integrates visual data with semantic descriptors. This framework will demonstrate how intermediate “concepts” (smoke, flames, etc.) can act as a powerful bridge between raw sensor data and predictive understanding. We anticipate this pipeline will serve as a blueprint for others working on interpretable deep learning in environmental monitoring.
* **Improved Forecasting Performance:** By fusing RGB and IR imagery with descriptors, our model is expected to achieve **higher accuracy in long-term fire spread predictions** compared to conventional image-only models. We will provide benchmark results showing the degree of improvement (e.g., percentage gain in IoU or extended prediction horizon) attributable to the additional modalities. This will quantitatively validate the hypothesis that more information (especially IR to see through smoke) yields better predictions of wildfire behavior.
* **Insight into Descriptor Usefulness:** Through ablation studies and comparisons, we will shed light on *which* descriptors are most useful for modeling wildfires. Perhaps we will find, for example, that a “smoke coverage” descriptor significantly boosts performance under certain conditions, or that “spread direction” descriptors help the model align its predictions with physical expectations. These insights contribute to the understanding of **explainable AI for wildfires**, identifying what high-level features a model should pay attention to. Even if some descriptors turn out to be redundant, knowing that is useful for designing simpler models.
* **Evaluation of Vision-Language Models in a Domain Context:** Our work will also provide a case study on using **LLMs (like GPT-4V and BLIP-2) for domain-specific image description**. We will report on how well these models, pre-trained on general data, handle wildfire imagery – information that is valuable to the broader community using foundation models for scientific applications. For instance, if GPT-4V can accurately identify smoke vs. cloud in an image, that is a notable capability. Conversely, if it struggles without fine-tuning, our analysis will highlight the gaps, possibly motivating further research into adapting VLMs for hazard monitoring.
* **Multimodal Fusion Strategies for Wildfires:** From our experiments, we will derive lessons on **effective fusion of RGB and thermal IR data** for computer vision tasks. We will document whether early fusion or late fusion worked better for fire forecasting, and how we integrated textual descriptions with visual features. This contributes to the broader literature on multimodal deep learning. Researchers in related domains (e.g., flood mapping with optical + radar imagery, or search-and-rescue with RGB + thermal cameras) can likely apply our fusion techniques and findings to their problems.
* **Open-Source Dataset Augmentation:** We plan to release the **labeled descriptor subset** of the Corsican dataset (the 100-200 images with our annotated descriptors) as an open resource to the community, if licensing permits. This will be a new dataset of wildfire images with human-described attributes, which can foster further work on explainable wildfire AI. Additionally, any code for fine-tuning BLIP-2 or our best-performing model will be made available to encourage replication and extension.

In conclusion, this research will advance the state-of-the-art in wildfire prediction by marrying the strengths of different data modalities and interpretability techniques. The outcome will not only be a high-performing forecasting model, but also a demonstration that **transparency and performance can go hand-in-hand** in deep learning for environmental safety. We expect this will pave the way for deploying AI systems that firefighters and planners can *trust* and *understand*, ultimately contributing to more effective wildfire management and risk mitigation. With wildfire frequency and intensity rising globally, such innovations are both timely and socially impactful.

**References:** (Selected inline citations) The Corsican Fire Database (CFDB) ([CORSICAN FIRE DATABASE | Projet Feux | Università di Corsica Pasquale Paoli | Université de Corse Pasquale Paoli](https://feuxdeforet.universita.corsica/article.php?id_art=2133&id_rub=572&id_menu=0&id_cat=0&id_site=33&lang=en#:~:text=This%20database%2C%20named%20the%20Corsican,range%20and%20near%20infrared%20range)) ([Deep Learning Approach for Wildland Fire Recognition Using RGB and Thermal Infrared Aerial Image](https://www.mdpi.com/2571-6255/7/10/343#:~:text=55,CrossRef)); Concept Bottleneck Models ([[2007.04612] Concept Bottleneck Models](https://arxiv.org/abs/2007.04612#:~:text=concept%20bottleneck%20models%20by%20editing,on%20concepts%20at%20test%20time)); BLIP-2 Vision-Language Model ([[2301.12597] BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models](https://arxiv.org/abs/2301.12597#:~:text=image%20encoder,can%20follow%20natural%20language%20instructions)); McKenzie et al. on infrared imagery penetrating smoke ([Seeing Through the Smokescreen: How Multi-Format Imagery Enables Effective Wildfire Response](https://resources.mckenzieintelligence.com/blog/seeing-through-the-smokescreen-how-multi-format-imagery-enables-effective-wildfire-response#:~:text=Shorter%20wavelengths%20present%20in%20visible,effectively%20rendering%20the%20smoke%20transparent)); Qayyum et al. on transformers for fire spread ([Machine Learning and Deep Learning for Wildfire Spread Prediction: A Review](https://www.mdpi.com/2571-6255/7/12/482#:~:text=)); CLIPScore for image caption evaluation ([CLIPScore: A Reference-free Evaluation Metric for Image Captioning - ACL Anthology](https://aclanthology.org/2021.emnlp-main.595/" \l ":~:text=image%2Bcaption%20pairs%20from%20the%20web%2C,Beyond%20literal%20description%20tasks)); EffcientNet CNN architecture ([EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks](https://www.researchgate.net/publication/333444574_EfficientNet_Rethinking_Model_Scaling_for_Convolutional_Neural_Networks" \l ":~:text=,)); etc. (Full bibliography will be provided separately.)