**Multi-Modal Wildfire Spread Forecasting with Descriptor-Guided Learning**

**Research Proposal**

**1. Introduction & Motivation**

Wildfire events pose a significant threat to both ecological systems and human communities. Despite recent advances in computer vision and machine learning, **accurately predicting wildfire behavior** remains difficult due to:

1. **Smoke Obscuration**: In standard **RGB** images, thick smoke blocks critical fire features. Infrared (IR) wavelengths, on the other hand, penetrate dense smoke, **rendering key hotspots visible** ([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))
2. **Black-Box Models**: Many state-of-the-art deep-learning systems provide **high accuracy** but minimal interpretability, lowering user trust—particularly in high-stakes domains like firefighting ([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))

**Descriptor-guided modeling** may help address these issues. Intermediate, human-readable attributes—such as “80% smoke coverage,” “flame front near the northern edge,” or “high thermal intensity”—can provide **explainable** signals for models and real-time decision-makers. Indeed, *concept bottleneck models* demonstrate that **predicting domain-relevant “concepts”** before final predictions can preserve predictive accuracy and boost interpretability ([*[2007.04612] Concept Bottleneck Models*](https://arxiv.org/abs/2007.04612)).

Furthermore, **vision-language models (VLMs)** such as **GPT-4 Vision** ([*[2303.08774] GPT-4 Technical Report*](https://arxiv.org/abs/2303.08774)) and **BLIP-2** ([*[2301.12597] BLIP-2*](https://arxiv.org/abs/2301.12597)) can automatically generate textual descriptions from images. By combining **RGB**, **IR**, and **descriptor-based outputs**, it becomes possible to develop a pipeline that (1) **sees** through smoke, (2) **learns** from interpretable concepts, and (3) **forecasts** wildfire spread with improved accuracy and transparency.

**2. Research Objectives**

1. **Multi-Modal Fire Forecasting**: Integrate **RGB, IR, and descriptor information** to predict fire spread more accurately than single-modality approaches.
2. **Descriptor Utility Analysis**: Compare the performance of (a) **image-only** models, (b) **descriptor-only** models, and (c) **fused** approaches, examining the effect of intermediate descriptors on predictive accuracy and interpretability.
3. **Vision-Language Descriptor Generation**: Evaluate **GPT-4 Vision**, **BLIP-2**, and other potential VLMs (e.g., **Google’s Gemini**) for **fire-related captioning** and descriptor creation.
4. **Temporal Modeling**: Implement **sequence-based** architectures (e.g., CNN-LSTM, Transformers) that handle multi-frame data and forecast how a fire evolves over time.
5. **Explainability**: Demonstrate how **descriptors** can provide reasoned explanations (e.g., “smoke drifting east → likely spread to the east next frame”) and validate whether it builds user trust.

**3. Data Source: Corsican Fire Database**

We will employ the **Corsican Fire Database (CFDB)** for this project:

* **RGB + IR Imagery**: CFDB provides aligned pairs of visible-spectrum and near-infrared frames.
* **Temporal Sequences**: Some images form multi-frame “video-like” sequences ideal for forecasting.
* **Annotations**: CFDB includes meta-data such as fire perimeter outlines, approximate smoke coverage, and other domain insights.

**3.1 Manual Descriptor Labels**

To train and validate descriptor generation, we will **manually label ~150 images** with wildfire-specific descriptors:

* Smoke coverage (e.g., 0–100%)
* Flame front visibility (yes/no)
* Spread direction (N, S, E, W)
* Embers or spotting presence
* Any notable color/thermal intensity remarks

These annotations will serve as **ground truth** for descriptor prediction tasks and help **evaluate** how accurate AI-generated descriptions are.

**4. Methodology**

Our approach comprises **descriptor generation**, **model baselines**, and **multi-modal fusion** aimed at robust wildfire forecasting.

**4.1 Descriptor Generation**

1. **Vision-Language Models**
   * **GPT-4 Vision** and **BLIP-2** are prompted with instructions (e.g., “Describe the wildfire scene: note smoke, flames, direction, color.”).
   * Descriptors are then automatically output for each CFDB image.
2. **Quality Evaluation**
   * We compare AI-generated descriptors against **human labels** via **BLEU**, **CLIPScore** and expert assessment (domain experts rating correctness and relevance).
3. **Descriptor Refinement**
   * If multiple VLMs are tested, we may keep only the best outputs or combine them in an ensemble.
   * For standardization, we will map free-form text to a concise set of domain descriptors (e.g., “smoke coverage = 60%, flame front = west edge”).

**4.2 Baseline Models**

1. **Image-Only**
   * **EfficientNet** (*[EfficientNet: Rethinking Model Scaling](https://arxiv.org/abs/1905.11946)*) for classification (e.g., predicting if fire spreads left vs. right).
   * **Vision Transformer (ViT)** ([*An Image is Worth 16x16 Words*](https://arxiv.org/abs/2010.11929)) for learning spatial features from single frames.
   * For temporal forecasting, wrap the CNN or ViT outputs in an **LSTM** (i.e., CNN-LSTM) to handle sequences.
2. **Descriptor-Only**
   * Fine-tune a **BERT**-based model to read textual descriptors over time (e.g., “time1: heavy smoke, time2: flame visible...”) and predict next-step outcomes (spread direction, extent, etc.).
   * This tests whether domain concepts alone suffice for meaningful predictions.

**4.3 Multi-Modal Fusion**

The **core** of our approach fuses **RGB, IR**, and **descriptors**. We explore two primary architectures:

1. **CNN-LSTM with Fusion**
   * **Early Fusion**: Concatenate IR and RGB channels (6 total) plus descriptor embeddings at the **input** stage, letting a single CNN process them together.
   * **Late Fusion**: Keep separate CNN streams for RGB and IR, plus a text encoder (e.g., BERT) for descriptors. Fuse learned features at a higher layer (e.g., fully connected layer before final output).
   * A **single LSTM** or **stacked LSTM** then models temporal evolution across frames.
2. **Transformer-Based Sequence Modeling**
   * Per time step, produce a **triplet of embeddings**: (1) RGB embedding, (2) IR embedding, (3) descriptor embedding.
   * Feed these combined embeddings (or separate tokens) into a **temporal Transformer** that leverages self-attention across frames and modalities.
   * Possibly incorporate an **attention-based interpretability** mechanism to highlight which descriptors or image regions drive the final predictions.

**4.4 Training Strategy**

* **Loss Functions**:
  + *Cross-Entropy* for classification tasks (e.g., “spread vs. no spread”), *segmentation losses* (IoU or Dice) for predicting fire masks.
* **Data Augmentation**:
  + Random flips, rotations, slight color jitter for RGB images, and potential synthetic IR generation if needed.
* **Transfer Learning**:
  + Since CFDB is moderate in size, we will use pre-trained models (e.g., **ImageNet**-trained EfficientNet/ViT, BERT from **Hugging Face**).
* **Validation**:
  + Cross-validate at the “fire event” level (training on certain fires, testing on unseen fires) to measure **generalization**.

**5. Evaluation Plan**

**5.1 Descriptor Quality**

1. **Automated Metrics**
   * **BLEU**, measuring n-gram overlap with ground-truth text.
   * **CLIPScore** to assess alignment between the generated descriptor and the underlying image.
2. **Human Review**
   * Domain experts (wildland firefighters or researchers) assess correctness of “smoke,” “flame front,” or “thermal intensity” descriptors.

**5.2 Static Classification**

* **Precision, Recall, F1**: For tasks like “Will the fire intensify next frame?”
* **Accuracy** can be misleading in imbalanced scenarios, so we emphasize **recall** (avoid missing potential spread) and **precision** (avoid false alarms).

**5.3 Spatial Prediction**

* **Intersection over Union (IoU)**: Overlap between predicted fire region and ground truth.
* **Hausdorff Distance**: Measures boundary alignment accuracy (lower distance = closer match).
* Evaluate these on **RGB-only vs. IR-only vs. fused** inputs, to see if IR and descriptors help.

**5.4 Temporal Spread Forecast**

* **Mean IoU over multiple frames**: Average how well the model predicts fire location for N steps into the future.
* **Multi-Step Drift**: Measure how quickly predictions diverge from reality with iterative forecasting.
* Possibly **Hausdorff** at final forecast frame to gauge worst-case boundary error.

**5.5 Interpretability Assessment**

* Examine **attention maps** or descriptor-based attention in Transformers to confirm that relevant features (e.g., “strong wind,” “dense smoke”) actually guide predictions.
* Conduct **expert user studies**: Compare descriptor-guided output (with textual explanation) vs. black-box output for trust and clarity.

**6. Preliminary Implementation Steps**

1. **Dataset Preparation**
   * Gather IR + RGB pairs from CFDB, select ~150 frames for descriptor labeling.
2. **Descriptor Generation Tests**
   * Prompt GPT-4 Vision and BLIP-2 on these 150 frames; measure BLEU, CLIPScore.
3. **Baseline Model Training**
   * Train EfficientNet (RGB-only, IR-only) and BERT (descriptors-only) to get initial classification/spread metrics.
4. **Fusion Model Development**
   * Implement CNN-LSTM (late vs. early fusion). Evaluate performance gains.
5. **Transformer Exploration**
   * Build a multi-modal Transformer for extended sequences, compare results to CNN-LSTM baselines.

**7. Future Extensions**

1. **Environmental Data Integration**: Add wind, humidity, topographical data for more holistic modeling
2. **Real-Time Edge Deployment**: Compress or distill the model so it can run on **drones** or **watchtowers** for live fire monitoring.
3. **Satellite and Aerial Imagery**: Apply pipeline to **VIIRS**, **MODIS**, or **drone-collected** large-scale IR images for real-world coverage.
4. **Synthetic IR**: If IR cameras are unavailable, investigate **GAN-based** or diffusion-based IR generation from RGB images.
5. **Generalization**: Evaluate on other wildfire datasets (North America, Australia) for broader adoption.

**8. Expected Contributions**

1. **Descriptor-Guided Pipeline**: A novel architecture that integrates **visual and textual** data for wildfire forecasting, moving beyond black-box methods.
2. **Improved Accuracy**: Empirical evidence that **IR + descriptors** outperform single-modal approaches in long-term fire spread predictions.
3. **Interpretability**: Demonstrate that textual descriptors can enhance human understanding of a model’s predictions and rationale.
4. **Vision-Language Model Benchmark**: Compare GPT-4 Vision, BLIP-2, etc. on **wildfire image captioning**, contributing to domain-specific performance insights.
5. **Open-Source Tools & Data**: Release annotated descriptor subsets of CFDB and relevant code (if licensing allows), encouraging replication and further research.

**9. Conclusion**

By **fusing RGB**, **IR**, and **descriptor-based features**, this proposal aims to establish a wildfire spread forecasting system that is both **accurate** and **transparent**. Through robust evaluation—ranging from descriptor quality checks to multi-step fire progression tests—this research will provide **new insights** into how interpretable ML can help tackle real-world challenges. Ultimately, our findings will not only advance **computer vision** and **NLP** integration but also offer a **practical tool** for first responders and environmental agencies combatting wildfires.