INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

Enhancing Vision Transformer with Vision language Model Text Embeddings for Robust Air Pollution Classification

Mohamed Mokhtar Ouardi[1], Dayang N. A. Jawawi[2*] & Farhan Mohamed[3]
Faculty of Computing, Faculty of Engineering
Universiti Teknologi Malaysia
81310, UTM Johor Bahru, Johor, Malaysia
Email: m.ouardi@graduate.utm.my[1]; dayang@utm.my[2]; farhan@utm.my[3]

Submitted: 30/9/2025. Revised edition: 14/4/2026. Accepted: 4/5/2026. Published online: 10/6/2026
DOI: https://doi.org/10.11113/ijic.v16n1.678

***Abstract*—Far from arguing the importance of ecological awareness in our current era it is however necessary to highlight the pressing global environmental challenges that are faced by our society. While air pollution is only a sector of concern it is essentially one of the most critical factors influencing human health and environmental sustainability. Under this premise, monitoring air quality is necessary. While the Air quality index has mostly been measured using Internet of Things (IoT) sensors, detecting visible air pollution has garnered interest due to its accessibility. However, the existing works based on vision only methods (CNNs, Vits) have shown limitations in capturing generalized correlations that are essential for a robust air pollution detection system. The proposed solution investigates the capability of semantic information to broaden the scope of features learned by the model. A Vision language Model (VLM) based text encoder with the objective of introducing knowledge anchors across any datasets. The model generates language tokens to guide a Vision transformer. The proposal also investigates tuning mechanisms for the VLM and image filtering for the input data. The key innovation targets a cross modal integration of vision transformers with a vision language model to create a few shots classification model for air pollution classification. The aim is to produce a model with flexible data integration and capable of leveraging visual and semantic correlations. The research demonstrated an improved generalization across a broad dataset variance. The model outperforms baseline CNN in accuracy when it comes to cross data implementation. The text tokens improve the attention maps focus on features that are exclusively related to air pollution. Ultimately, looking into the impact of a flexible framework for environmental monitoring is crucial, not only for air pollution detection but also for broader environmental challenges.**

***Keywords*—Vision Language Models, Multimodal Learning, Computer Vision, Environmental Monitoring**

I. INTRODUCTION

Citing the American Lung Association to highlight the air pollution trends across the last 5 years which displays 112 countries that earned an F grade for unhealthy spikes in particulate matter air pollution. As a matter of fact, the 2024 State of the Air report demonstrates the highest spikes of fine particulate air pollution. In fact, unexpectedly this follows years of better air quality index and a reversal of the pollution trend. The establishment of the Clean Air Act has played a major role in slowing down the negative trend. However, with the recent wildfires which are the main contributing factor to the increasing number of days and places with unhealthy levels of particle pollution in recent years [2], [3]. The recent report designated the highest number of days with severity of pollution exceeding the hazardous threshold. Between 2020 and 2023, upward of 32 million people across about 60 countries have been listed under high-risk pollution zones. Looking into the existing health reports, a pattern of health risk emerges, showing an obvious correlation between prolonged exposure to polluted air and respiratory diseases [1]. Consequently, the escalating trend underlines the urgent requirement for more accessible and adaptive monitoring methods that are capable of localizing high risk pollution areas.

Given this context, it is necessary to address the relevance of real time vision pollution detection monitoring systems. The first factor is accessibility, the cost effectiveness of a camera-based pollution detection is nearly unmatched. Not only can it leverage the existing infrastructure but it is compatible with most existing models on the market. Similarly, when it comes to scalability the installation of new detection nodes does require little to none expert intervention. When compared to

IoT devices, a dense network requires large operation for implementation and calibration which in rural areas where the technology is most needed becomes a hurdle to the adoption of the monitoring systems [4]. Equally as important, is the area of coverage which cameras offer a wide field of views that depending on the location can span kilometers. Even though the argumentation does not undermine the reliability of IoT sensors, the integration of vision-based methods with an IoT network can become a complementary framework that strengthens the monitoring capabilities [5].

To emphasize the impact of this research, it is necessary to highlight the limitations of existing methods. Therefore, hardware-based solutions struggle in a few domains. Starting with the cost behind the installation of an air quality monitoring station which in low-income regions becomes a major obstacle to the deployment of the technology. Following, the sparse spatial distribution may lead to blind zones across the network which lack supervision. Additionally, vulnerability of the sensors to noise and external disruptions, create reliability challenges especially when it comes to the longevity of the system. And ultimately, from the user side the blurry interpretability of the sensors. Indeed, the method tends to fail to transform the numerical readings into concise human friendly interpretation.

When it comes to vision-based methods, the challenge of existing works is generalization. The lack of domain flexibility and the susceptibility to overfitting is a fundamental hurdle when it comes to integration in the real-world domain. The range of difficulty between a controlled environment and a real-time highly variable context is substantial, as models trained under a fixed setup tend to encounter critical failures when exposed to real world factors such as weather and lighting [7], [8]. This ultimately reveals the requirement of vast amounts of data just to attenuate the risks of overfitting. Fundamentally, every single vision-based method is striving to achieve robustness and interpretability.

With these points in mind, the study focus shifts to multimodal approaches. The concept of combining semantic information and visual cues relies on the complementary of the information. Defining air pollution in a complex environment exclusively with vision data can be challenging. In fact, semantic enrichment helps guide the vision tokens towards what has the highest rate of learning convergence for air pollution classification. The improved generalization that is induced from the novel data comes from the high-level features that remain stable across diverse contexts which not only help with flexibility but also can serve as a token for interpretability [12], [13], [15]. For instance, in the case of vision-based models it is hard to understand the reasoning behind the prediction fundamentally facing a black box that can only be modified by training the model. In the case of the proposed approach the text tokens can also serve as an indicator to the user on the behavior guiding the model to the prediction.

Henceforth, the Enhanced Vision Transformer with VLM Text Embeddings for Robust Air Pollution Detection (ETE-ViT) is proposed. The ETE-ViT framework leverages text embedding generated from vision language models to enrich the input of a vision transformer. Furthermore, it caters the input data to optimize the features relevant to air pollution classification. This approach handles the region of interest through the implementation of an object detection model (Yolov12) designed to perform foreground background segmentation [21]. Following are a series of nodes that are responsible for guiding VLM prediction and enhancing the input image clarity by minimizing the real-world noise. The aim is to attain a robust model capable of handling adverse and variable conditions. ETE-ViT is built to mitigate the influence of domain shift of the accuracy results. The approach addresses the problems of poor generalization in vision only pollution detection. Furthermore, this framework is modular and extensible both across different use cases but also through the integrability of external data nodes such as hardware sensors. Accordingly, the existence of semantic text provides a direct interpretability of the model prediction which is important when looking for the adoption of technology. The additional capability of reading the initial reasoning of the model does provide trustworthiness to the end user. Ultimately, ETE-ViT aims to contribute to the interpretability, scalability and generalization of vision-based models. In this specific research, air pollution classification is the chosen use case for the environmental relevance and the challenging problem.

In summary, to clarify the research gaps prior work rely heavily on CNNs and ViTs which are constrained by purely vision input. These methods excel in controlled environments but fail drastically in domain shifts such as different lightning, weather and even regions [7], [9], [10]. This comes back to the challenge of generalization. Furthermore, hardware focused approaches face the hurdle of accessibility and scalability. The solution has a relatively limited range of action. They require specific infrastructure to be optimal. In the context of air pollution there is an observable correlation between the risk zones and their financial capacity which fundamentally requires methods that are highly accessible with little to no expertise requirement for the deployment. This also extends on the concept of interpretability, numerical outputs or raw inference results may limit the end user familiarity to the output which by leveraging the semantic text this can be improved.

The research objective is to design a flexible VLM-guided framework to enhance vision transformers prediction for air pollution classification. This occurs through a highly generalized model that can better interpret data heterogeneously while being a modular design ready for other multimodal integration in the future.

Definitely, on air pollution as a case study focusing on classifying the risk index of the supervised area. Demonstrate that multimodal approaches can compete with hardware-based methods while being both more accessible and having a bigger range of supervision. The potential of real-world application is an essential factor in the research scope. The design is built around the idea of public health through offering interpretable real time environmental clarifications while also being a lookout system for high-risk accidents such as wildfires. Ultimately, the use case offers multiple areas of implementation ranging from forest monitoring up to farm supervision and even personal home air supervision. For this specific research, it is focused on urban air pollution.

The following sections of the paper are structured as follows methodology, result and discussion and a conclusion.

II. METHODOLOGY

The methodology section outlines the research design used through the development of the ETE-ViT Framework. Given the objective of developing a framework that is scaleable, interpretable and accessible catered to air pollution classification, a modular approach is adopted. The framework consists of 5 modules. Region of Interest Filtering, CLAHE & Dehazing and Edge optimization under the umbrella of Input preprocessing for optimal in context performance. Followed by the Vision Language Model (Qwen 2.5) which is the text encoder. Additionally, connected to the VLM is a Guidance Module that encapsulates both the prompt augmentations, enrichment resources and hardware data when available. Ultimately, the VLM output is tokenized in order to be integrated in the Vision Transformer. This modular approach provides a flexible structure that can be enhanced in future works. The main contribution of this proposal lies in the guidance module. The multimodal framework introduces a novel integration of visual, textual and contextual data streams. This converts the baseline vision transformer into an explainable AI model with a high-level interpretability and automated reasoning for augmentation tasks. In fact, the guidance module refines prompts and regulates information flow between data nodes while allowing the pipeline to be dynamic and modular. As a summary, this section covers each module design along with the training setup, the data used, the evaluation metrics and the implementation details.

A. *Overview of the Framework*

As demonstrated in Fig. 1 the framework is subdivided into 5 main modules. A foreground / region of interest filtering which consists of an Yolov12 foreground background segmentation model along with a thresholding algorithm. A preprocessing module consisting of 3 algorithms Contrast Limited Adaptive Histogram Equalization (CLAHE), Dehazing and Edge sharpness optimization. A VLM module composed of Qwen 2.5 playing the role of the text encoder. A guidance module which is multiplexing prompt augmentations, semantic priors and auxiliary sensor signals to refine the VLM's reasoning. Finally, the Vision transformer module which consists of the model backbone along with the tokenization and the embedding layer for the image and text tokens. The integrated pipeline produces a generalized representation of the air quality indicators and features within the input data. Ultimately, the framework role is to forward the reasoning to the end user along with a classification of the air quality index and its prediction confidence.

Unlike existing works, the proposed framework leverages multimodal data which effectively support cross domain learning. The integration of visual data and language models behave as a conduit toward few shot learning. The pipeline capitalizes on the language tokens specifically when enriched with hardware sensors to guide the inference output even when lacking familiarity with the dataset [12], [16]. This inherently offers the model the ability to use these contextual anchors as regularization signals when performing fine tuning.

The guidance module does offer a feedback loop for 2 objectives. Leveraging a form of behavior cloning, it applies the end user's comment on the output to update the prompt augmentation. And secondly, it translates the output prediction into a semantically meaningful concise clarification of the result which enables end users to adopt the approach with little to no prior expertise of the domain.

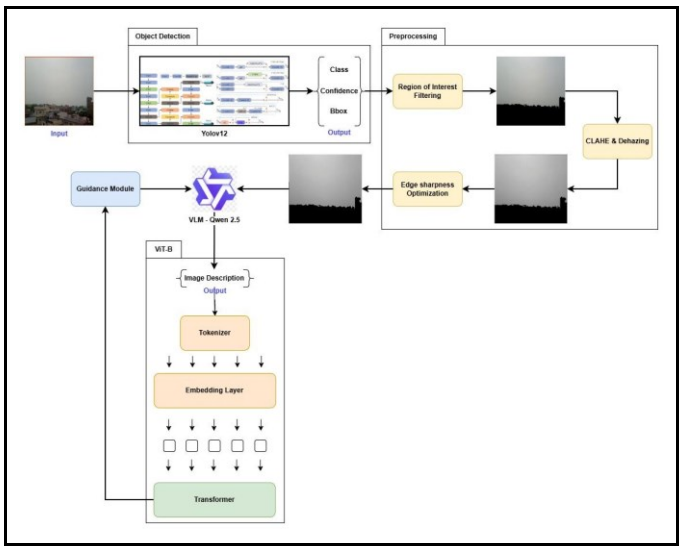


Fig. 1. ETE-ViT framework

B. *Region of Interest Filtering*

The reasoning behind the introduction of the Region of Interest Filtering comes from the amount of irrelevant elements on raw environmental images. Often buildings, cars, people do not contribute with valuable features for adequate correlation. Considering the highly unpredictable variance of data in the use case of air pollution detection it is necessary to localize the region of interest beforehand. By smoothing the amount of variance in data, the produced model has a more refined understanding of Air quality index classification.
Yolov12 is a state-of-the-art segmentation model which is highly compatible for the task of foreground background masking. The forward pass identifies the regions of interest such as the sky, smoke clouds, haze. This approach allows transient detections to be filtered out [21].

A thresholding algorithm is applied to the detection based on the confidence score giving trainers the ability to adjust the rigidity of the RoI filter. This ensures that the filter is tunable according to the camera setup.

The output is a masked version of the input minimizing the irrelevant noise in the image input. This filtered image ideally only contains regions with high relevance to air pollution. Furthermore, it demonstrates an improved signal to noise ratio in the dataset. In fact, between enhancing the accuracy by focusing on the semantically relevant areas and reducing the false positives caused by the foreground noise the RoI filtering demonstrates its role as a critical presprocessing stage. Ultimately, by narrowing down the input information to pollution relevant regions, it improves detection robustness and VLM description quality.

C. *Image Preprocessing*

Continuing on the purpose of improving the quality of the input dataset. This research proposes to implement 3 algorithms with the aim of achieving a cleaner and more catered dataset. The 3 algorithms are performed online during both training and inference.

Contrast Limited Adaptive Histogram Equalization (CLAHE) with the reasoning of improving local contrast and visibility in low light conditions. CLAHE is optimal to ensure a balanced contrast across different contexts while preventing over amplification of noise.

$$I'(x,y) = \frac{CDF_{clip}(I(x,y)) - \min(CDF_{clip})}{\max(CDF_{clip}) - \min(CDF_{clip})} \times (L-1) \quad (1)$$

Equation (1) rescales pixel intensities according to the clipped cumulative distribution function. The local histogram tiles are averaged in order to prevent noise accumulation in the augmented image. Additionally, the normalization is implemented to tighten the range of intensity to a maximum of (L-1). This function produces an improved contrast especially in images with low exposure or blurry edges. The CLAHE is a localized algorithm therefore the filter is not applied to the entire image but for the specific areas requiring the enhancement. The algorithm ensures that the image contrast does not over blow in the broad scope of the image.

Dehazing function is required to remove scattering effects that can either be generated through the CLAHE process or existing due to the blurry nature of air pollution (e.g. smoke, smog, clouds). The objective is to find the optimal radiance of the affected regions which improves the visibility and contrast between merged elements such as the sky and smoke. This is an important algorithm to pop important features that would be otherwise lost through particle scattering [19].

$$J(x) = \frac{I(x) - A}{\max(t(x), t_0)} + A \quad (2)$$

Equation (2) demonstrates J(x) which is the optimal radiance that is looked for. This is obtained by measuring I(x) the input image and then compensating for the light A and the transmission map t(x). In principle, the function allows to recover the unblurry representation of the scene. This ensures that the image has distinguishing features between the background and the pollution relevant elements. This is essential to localize smoke, clouds and smog from the sky. This ultimately helps to highlight the relevant features for the use case.

Gradient based Edge Sharpness is a structural detail improvement. Following the dehazing process most features have distinguished layers. However, there is still a range of occlusion and slight visibility limitation. In order to improve the clarity of the element's boundaries and generate sharper features a sober operator function is applied [20].

$$S(x,y) = \sqrt{G_x^2 + G_y^2}, \quad S_{avg} = \frac{1}{N}\sum_{x,y} S(x,y) \quad (3)$$

In Equation (3) the gradient magnitude S(x, y) calculates the intensity variance through both the vertical and horizontal axes (G_x) and (G_y). The high gradient values are representatives of edges in the image. Following, the algorithm measures the average sharpness score S_{avg} which offers the global edge clarity score. This is used to evaluate the overall sharpness across the image.

$$I'(x,y) = I(x,y) + \beta \cdot S(x,y) \quad (4)$$

Once the average sharpness is calculated equation (4) High Boost Gradient Enhancement is used to enhance the score by an optimal magnitude. β is a scaling factor for the gradient reinforcement increasing the difference between adjacent pixels along edges. This is measured by dividing a constant factor by S_{avg} which results in blurrier images getting a stronger boost. This is essential to avoid the trade off between sharpness gain and noise amplification.

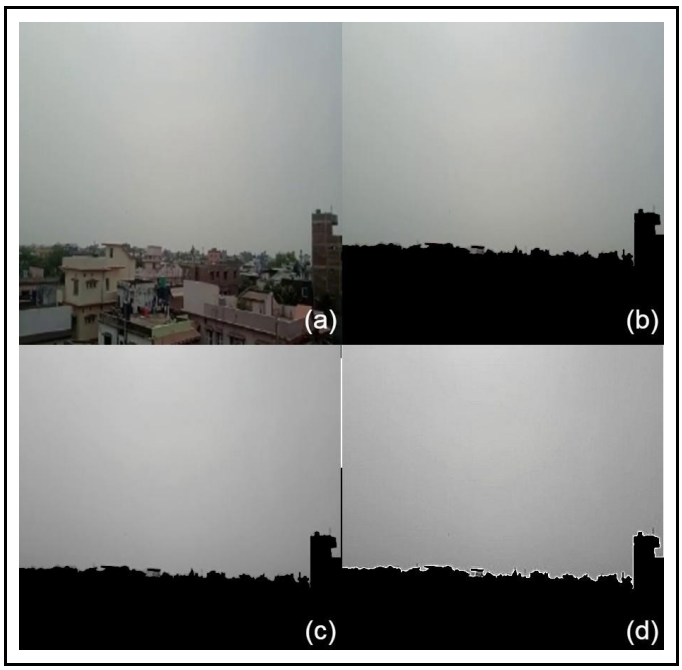


Fig. 2. Preprocessing sample (a) raw image (b) RoI Filter (c) CLAHE & Dehazing (d) Edge sharpness

Fig. 2 is a representation of the image transformation across each step of the preprocessing. This phase is essential to optimize the quality of the data procured to the model. Clean, high contrast data is directly proportional to the semantic richness of the image. The combination of the 3 algorithms aims to enhance the contrast, the visibility and the structural clarity of the features relevant to air pollution classification. In the context of a real-world environment, it is difficult to control the noise and variability occurring due to the external factors. Therefore, this preprocessing stage ensures a robust baseline for a reliable downstream inference.

D. Vision Language Model (Text Encoder)

Qwen 2.5 is an accessible open source and state of the art vision language model. The VLM is selected for its multimodal reasoning capabilities across both text and image domains. The convenient aspect of an open ecosystem represents a strong attribute for research purposes. Leveraging Qwen 2.5 as the VLM of the pipeline ensures a reliable image contextualization agent. The VLM is in charge of 2 main tasks, the first focus is to describe the image into semantic text. This is basically translating the image content into textual description that contains uniform context across any datasets. Furthermore, the VLM has the ability to concatenate additional knowledge into the generated tokenized description. This can include reasoning with a baseline knowledge such as specific research papers with adequate vocabulary for the inference or even hardware sensor data. As an example, Fig. 3 shows samples of tokenized textual output generated by the VLM. These demo the difference between raw VLM description on Fig. 2 sample (a) and when hardware sensors are included (b).

```
["the", "sky", "is", "uniformly", "gray", "with", "haze", ",",
"showing", "low", "visibility", "and", "weak", "atmospheric", "clarity"]
```
(a)

```
["the", "sky", "appears", "hazy", "and", "gray", "with", "low",
"clarity", ".", "sensor", "data", ":", "pm2.5", "95", "µg/m³", ",",
"pm10", "120", "µg/m³", ",", "co₂", "680", "ppm", ",", "humidity", "72",
"percent"]
```
(b)

```
["the", "sky", "is", "clear", "and", "bright", "with", "high",
"visibility", ".", "sensor", "data", ":", "pm2.5", "8", "µg/m³", ",",
"pm10", "15", "µg/m³", ",", "co₂", "410", "ppm", ",", "humidity", "45",
"percent"]
```
(c)

Fig. 3 sample VLM output (a) using Fig. 2 sample without hardware sensor measurement (b) using Fig. 2 sample with hardware sensor measurement © using a good sample with hardware sensor measurement

Qwen 2.5 is tuned using augmented prompts that dictate its behavior and the range of action. This is a convenient method to allow on the fly personalization by the end user without the intervention of the trainer.

On a more technical point, the text encoder embeds the output of the VLM into the tokenized patches ready for integration in the ViT. This module dedicates 256 patches to the VLM contextual text. This embedding approach is great modular convergence. The controlled allocation is great for balancing the semantic patches with the visual tokens in order not to collapse the model weights. The text context represents 11% of the total patches. And in order to maximize the value of the text tokens, an attention scaling is implemented on the deep attention layers of the model. This method is important for the integration of text patches without disrupting the pretrained architecture of the Vision transformer [12], [13].

E. Vision Transformer (Vision Encoder)

The vision transformer represents the core of the framework. It is utilized for visual feature extraction by encoding raw image patches into dense embedding that can be aligned with the text tokens. The ViT is a great multimodal architecture that can leverage both data sources and still be able to provide spatial attention mechanisms to capture pollution relevant features. This is achieved by controlling the text patch influence through the initial attention layers by scaling it to less than 5% of the token structure [7].

The model patch tokenization follows a standard of 640x640 input image divided into patches with window size 14x14 with a 1x1 padding. Each patch is flattened and linearly projected into an embedding vector. In order to maintain the spatial grid, the text tokens are fed in an exclusive grid parallel to the visual patches which ensures they do not override the positional information of the image. This is a good approach in order to separate the alignment and maintain the relevance of the visual patches while maximising the semantic grounding of the text patches. The joint attention layers get to learn a cross correlation to interpret spatial information with an enriched data context as demonstrated in Fig. 4.

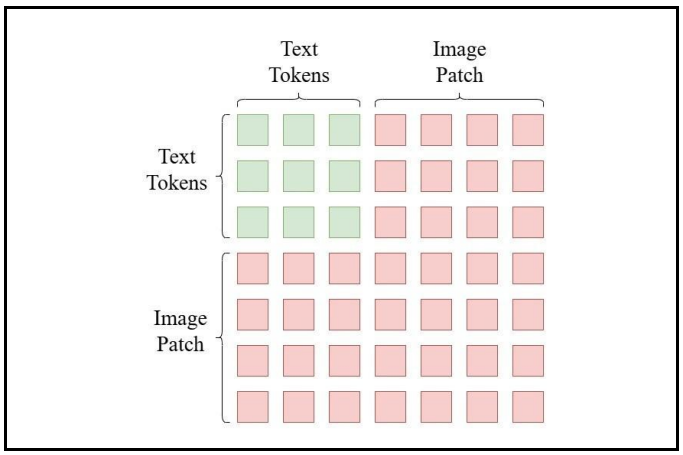


Fig. 4. Patch Tokenization

The ViT integrates multi head self-attention layers to capture dependencies across the full range of the features which is essential to correlate pollution indicators. In this context, the text tokens help align the features and distinguish between normal atmospheric patterns and pollution signals.

Additionally, similarly to a title assigned to an image the multimodal attention layer allows the vision tokens to query the contextual information provided by the text tokens.

The model is built upon ViT-B which is a stable pretrained weight with strong visual representations. This structure can be fine-tuned in the multimodal setup with minimal influence on the pretrained vision weights. This is achieved by preserving most of the backbone's architecture identical to vision only models and implementing the multimodal attention to deep layers. The model head consists of a Positional Encoding Transformer Encoder Classification (MLP) that outputs Air quality index classification and a confidence score [11], [14].

F. Guidance Module

The guidance Module is the control center for the text data of the ETE-ViT framework. The module integrates multi source data to provide a coherent control prompt to the VLM. The approach aims to enhance the flexibility of the pipeline by maximizing the controllable nodes. Running in parallel with ViT but without the real time constraints allows the integration of a deep reasoning LLM that can simulate an expert agent controlling the VLM and interacting with the end user [22], [23]. In essence, the guidance module consists of 5 elements as shown in Fig. 5.

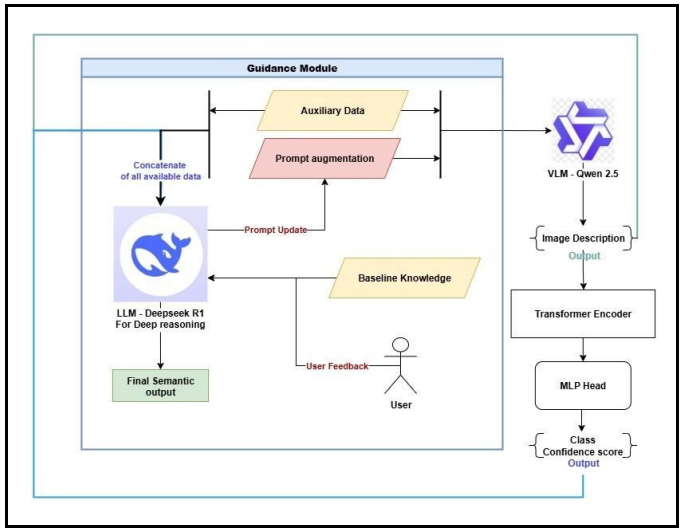


Fig. 5. Guidance module

An auxiliary data submodule that serves as input node for any form of relevant data for the air quality classification. This allows the framework to easily integrate hardware sensor data when available. The additional modular data inputs are used to improve the predicted tokens as demonstrated in Fig. 3. Further, reducing the reliance on the VLM's training and increasing the specialized knowledge available for the classification and final semantic output.

The prompt augment represents the control prompts that are fed to the VLM during the inference. These are instructions on the format, the size and the content of the semantic text contextualization. They are a great method to update the behavior of the VLM with no training requirements. This can be controlled by an LLM agent that can maintain the correct structure while using the end users' feedback to improve the VLM's output.

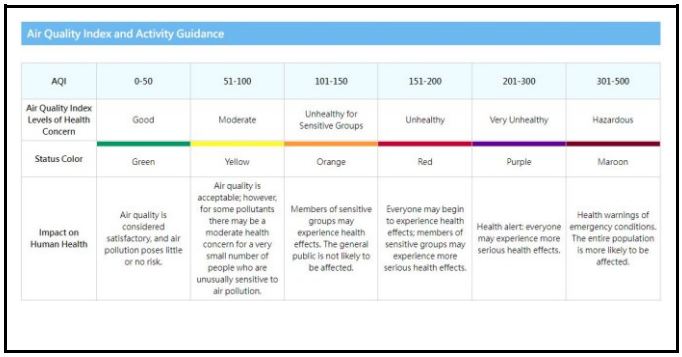
The baseline knowledge is a module dedicated to the LLM; it contains research papers and relevant vocabulary necessary to have optimal reasoning for the context of air pollution. This also ensures that the model won't hallucinate when answering edge case requests.

The user feedback is an accessibility module that aims at giving the end user some control over the classification performance and the final output. This is specifically relevant in the context of the deployment of the framework.

The Large Language Model is essentially the brain of the guidance module. The VLM excels at perception but fails when it comes to reasoning chains. This is where the LLM bridges the gap. It allows high level reasoning in order to integrate the inputs into a well formatted control prompt for the VLM. In this research, Deepseek R1 has been chosen due to its open-source architecture and accessibility. The model is a very performant reasoning agent while not requiring excessive resources to implement. The LLM is mainly used for the tasks within the guidance module. First of all, it represents the communication node with the user. This is mandatory to translate the user's unstructured feedback into the required format for the VLM while making sure to not collapse the functionality of the framework by giving the end user too much control over the control prompt. More importantly the LLM is used for interpretability. Instead of displaying the prediction output of the ViT which are only a classification along with a confidence score the LLM concatenate the inference output, the prior VLM contextualization and the auxiliary data into a concise, human friendly description.

G. Training Setup

To begin with the dataset used for this research is sourced from Eff-AQI: An Efficient CNN-Based Model for Air Pollution Estimation: A Study Case in India [6]. The dataset consists of an image dataset labeled according to the measured Air quality index and classified according to the impact on human health as demonstrated in Fig. 6.



Air Quality Index and Activity Guidance

AQI	0-50	51-100	101-150	151-200	201-300	301-500
Air Quality Index Levels of Health Concern	Good	Moderate	Unhealthy for Sensitive Groups	Unhealthy	Very Unhealthy	Hazardous
Status Color	Green	Yellow	Orange	Red	Purple	Maroon
Impact on Human Health	Air quality is considered satisfactory, and air pollution poses little or no risk.	Air quality is acceptable; however, for some pollutants there may be a moderate health concern for a very small number of people who are unusually sensitive to air pollution.	Members of sensitive groups may experience health effects. The general public is not likely to be affected.	Everyone may begin to experience health effects; members of sensitive groups may experience more serious health effects.	Health alert: everyone may experience more serious health effects.	Health warnings of emergency conditions. The entire population is more likely to be affected.

Fig. 6. Dataset classification format [6]

The dataset is collected across 2 countries Nepal and India with a total of 8 cities investigated. The data has been labeled according to these measurements of PM2.5, PM10, NO2, SO2, CO. It contains 12240 images distributed across 6 AQI classes with a ratio per class as shown in Fig. 7.

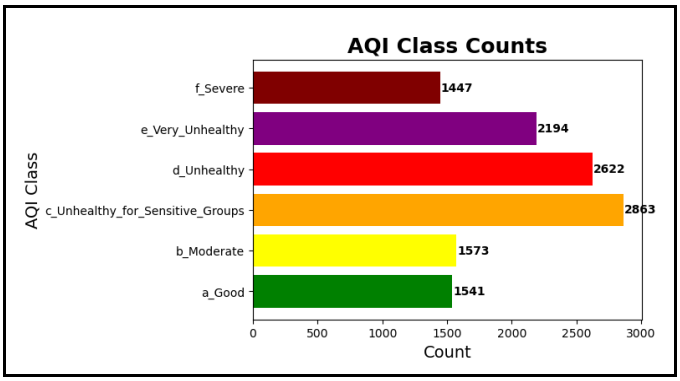


Fig. 7. AQI class distribution [6]

For the training the data split strategy selected consists of performing class distribution balance which ends up being 1447 images per class for a total of 8682 total images which are split in the ratio 70% training 20% validation 10% testing. Whenever it was possible the 10% testing split was exclusively extracted from the data collected in Nepal in order to have the training dataset being drastically different to the testing dataset [4].

Certainly, for more details on the training the ViT utilizes a symmetric cross-entropy loss function. The image and text embedding are L2-normalized and project into a common space. The loss averages cross-entropy over rows and cross-entropy over columns. Furthermore, the optimizer used is Adam with decoupled weight decay regularization and weight decay equal to 1e-2. The learning rate schedule follows a cosine decay with a peak value equal to 4×10^{-4}. The batch size is equal to 8 with 130 epochs in total. The implemented ViT-B architecture contains 12 layers with a hidden dimension of 768. This is trained on a NVIDIA T4 GPU with 16 GB VRAM.

The framework is evaluated qualitatively on the full pipeline output and quantitatively on the ViT classification performance. The ViT is compared to existing vision only models that are trained on the same dataset. The evaluation metrics are accuracy to get the overall ratio of correct predictions, precision to get the proportion of positive predictions that are correct, recall to measure all the actual positive instances and Specificity to measure all the rate of true negatives.

III. RESULT AND DISCUSSION

As a brief recall on the objectives of this research, ETE-ViT is a framework based on a multimodel transformer combining text and vision embedding. The proposed approach leverages a VLM (Qwen 2.5) to contextualize input images into semantic description of the scenery specifically targeting relevant information about air pollution. The pipeline implements an LLM (DeepSeek R1) for deep reasoning and modular guidance. The aim of ETE-ViT is to improve accessibility, interpretability and generalization. Providing a model pipeline that offers both robust and ready for real world implementation. And this goes from the accuracy of the model to the end user experience.

With this context reestablished, the results consist of quantitative and qualitative results. The testing is sectioned by modules as follows VLM, LLM, Preprocessing, ETE-ViT and Final Semantic Output. Each testing includes comparison with baseline models trained on the same dataset which as a reminder is 8682 images of urban images labeled across 6 classes according to their Air Quality Index relative to human health risk.

Starting with Fig. 8 representing, average cosine similarity between image and text embedding for different VLMs. The graph demonstrates how each model text output corresponds to the actual image scenery. This aims to evaluate the semantic understanding of the selected VLM on the use case of Air pollution and specifically the preprocessed format passed to the model. The average cosine similarity reflects the relevance of the description to the image content. The higher score means better comprehension of the environmental context. Qwen 2.5 is compared to 3 other VLMs (CLIP, BLIP-2, LLaVa-1.6) where the selected model consistently achieves the highest similarity score across the testing dataset [12], [13], [14], [16]. The correlation between ambiguous class scores (Moderate and Unhealthy) and the gap between model performance demonstrate the model's reliability in edge cases. In this research's use case Qwen 2.5 demonstrates a superior semantic text output with higher quality description for the ViT text tokens embedding. This enforces the role of VLM in improving the reliability of the overall framework output by providing consistent context across data variance.

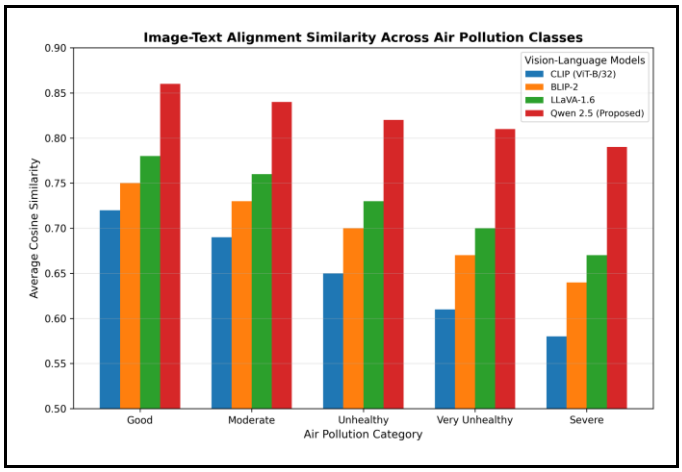


Fig. 8. Image-Text Alignment Similarity Across Air Pollution Classes

Fig. 9 is a prompt compliance diagram, which monitors the capability of a model to follow predetermined instructions. The model output is tested on the capability to follow the demanded reasoning form and output requirements. Both tests are aimed at testing the model combination of visual cues and contextual reasoning to generate an interpretable output for this study use case [24]. The tasks are designed to simulate the air pollution image interpretation under the context of this research. Prompt compliance is computed as the proportion of outputs to meet the expected response format, which includes all the requested reasoning components. Increased compliance means that the model can effectively align with task related prompts, a critical aspect in environmental monitoring contexts where reliability is essential. The findings indicate that the models that have high reasoning abilities will be more inclined to give consistent

responses that are consistent with the structured prompts employed in the proposed framework. This is of particular significance to multimodal environmental analysis where visual interpretation should be supported by consistent reasoning. In general, low compliance is a relevant indicator of hallucination for multimodal pipelines. In this case, DeepSeek-R1 displays the highest alignment with this research use case.

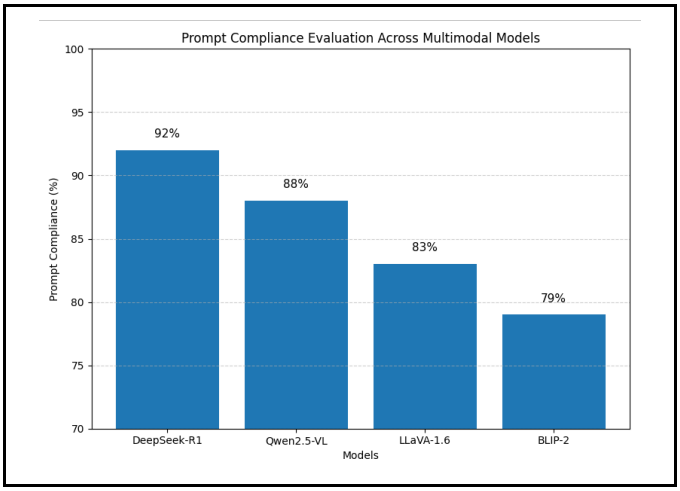


Fig. 9. Prompt Compliance

Table I focuses on the preprocessing stage of this framework.

TABLE I. ABLATION ON PREPROCESSING

Variant	Accuracy (%)	macroF1 (%)	Δ vs Full (%)
Full	84.0	0.823	–
No RoI	78.6	0.769	↓ 5.4
No CLAHE	81.7	0.801	↓ 2.3
No Dehaze	79.5	0.773	↓ 4.5
No Edge	82.9	0.813	↓ 1.1
Only RoI	80.4	0.782	↓ 3.6
Raw	76.8	0.758	↓ 7.2

The Ablation experimentation investigates the impact of the preprocessing components on the classification performance of ETE-ViT. The "Full" variant refers to all the preprocessing modules enabled; this is the evaluation baseline and it achieves 84% accuracy with a macro-F1 score of 0.823. Notably, macro-F1 is the function that computes f1 for each label, and returns the average which is a good representation on the influence of the preprocessing on individual classes. A score of 1 would mean that the accuracy across every label is equal. The table shows that both RoI and Dehazing are important preprocessing functions that drastically influence the model accuracy. Indeed, when RoI is removed the accuracy drops by 5.4% and without dehazing the accuracy drops by 4.5%. Ultimately, ablation experimentation demonstrates the value of all implemented functions. However, it is in the realm of optimization that Edge sharpness and CLAHE find their value with both having about 1.1% and 2.3% influence on the performance. Ultimately, the preprocessing module displays a relevant boost to the model output reaching 7.2% drop when disabled.

Following Table II is the quantitative benchmark of the ETE-ViT model compared to baseline transformers architecture commonly used for visual classification. ETE-ViT scores 84% accuracy displaying an overall prediction quality across all air pollution classes. This reflects a good generalization to unseen data. This outperforms and improves on the baseline architecture with Swin-T being competitive with 80.9% accuracy just as a vision only model [9].

The proposed framework reaches a precision of 83.4% which is an indicator of reliability of position prediction. This is an important representation of the low false positive rate generated by the model.

Furthermore, the recall measured at 82.1% shows a good generalization across all the classes with little overlooked edge cases. A strong recall represents a robust classifier capable of near optimal feature extraction of environmental patterns.
Finally, specifically the rate of false pollution predictions which is in the use case can be a critical failure if an environment is falsely labeled as being a high-risk area. Indeed, the metric is essential to avoid false alarms and is a relevant argument to exhibit the trustworthiness of the prediction. ETE-Vit scores 80.6% in this metric which is a strong result but it falls behind the baseline models such as Swin-T scoring 84.1% [9]. This is a point of concern, with roots embedded in the fragility of the boundaries between visually similar but semantically distinct air conditions.

Overall, ETE-ViT in the use case of Air pollution classification shows a competitive edge both in quantitative results but also in regards to the qualitative features that are catered to deployment and usability in real world settings.

These results validate the integration of VLM derived text embedding as a beneficial guide for multimodal vision transformer learning.

TABLE II. COMPARISON BETWEEN ETE-VIT AND BASELINE TRANSFORMERS

Variant	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)
ViT-B/16	77.8	76.9	74.8	81.2
DeiT-S	79.6	78.8	77.5	83.0
Swin-T	80.9	80.2	79.6	84.1
ETE-ViT	84.0	83.4	82.1	80.6

To support the importance of the ability of differentiating polluted from clean air across varying classification thresholds Fig. 10 demonstrates a Receiver Operating Characteristic curve (ROC) curve. This evaluation is performed by grouping the classes into 2 categories: "clean" (composed of labels "good" and "moderate") and "polluted" (composed of the 4 other classes). This is graphical representation that shows the performance of a binary classification is fundamental to show the reliability of the model in real world context with potential reproduction for wrong prediction. Across all tested models the

ROC displays a strong difference between positive and negative classes which validates the trustworthiness of the ViT classifiers.

ETE-ViT does achieve a better balance between detecting polluted conditions and avoiding false alarms compared to the baselines; however, the consistent performance stability seems to be an inherited feature of the vision transformer models.

Altogether, the ROC curve only reinforces that the framework not only performs in static accuracy metrics but also demonstrates robust prediction across the full range of detection thresholds. An indicator of the reliability of the transformer and adaptability to the use case.

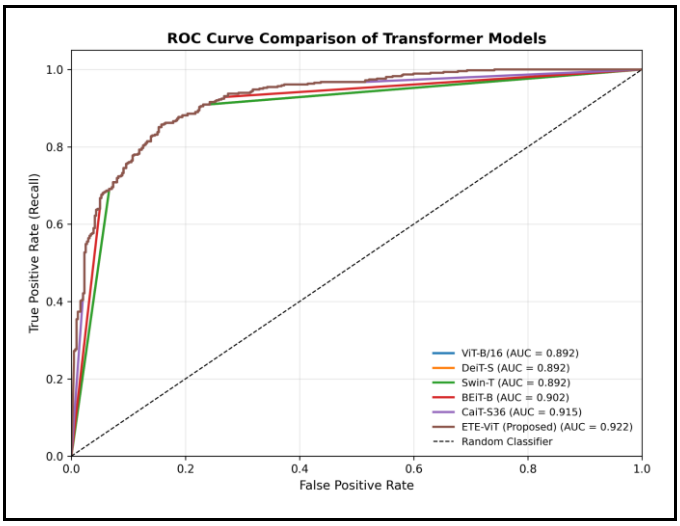


Fig. 10. ROC Curve Comparison of Transformer Models

In order to visualize the per class prediction accuracy of ETE-ViT Fig. 11 shows a confusion matrix. Notably, the diagonal density displays the model capability of differentiating between most pollution level with relatively low cross class confusion. Additionally, the high accuracy in "Good" and "Severe" classes shows a clear break between the two extremes of the learning curve.

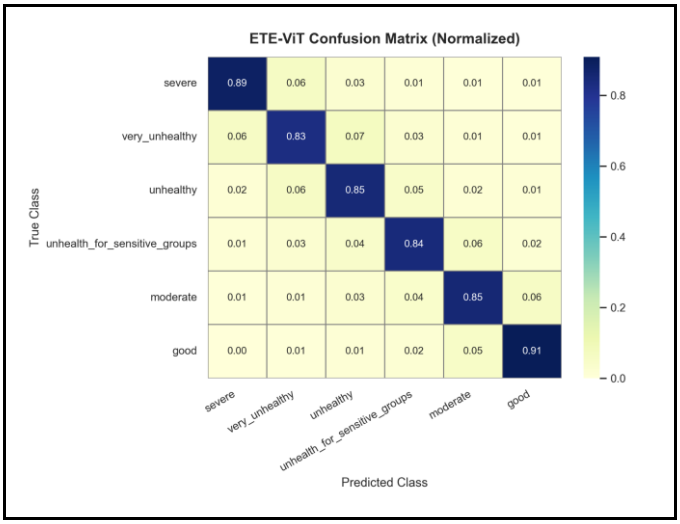


Fig. 11. ETE-ViT Confusion Matrix

The minor overlap among the mid-level classes is expected. In fact, the lower accuracy is observed in highly ambiguous cases with rarely enough strong distributions to influence the features towards the expected ground truth. Ordinarily, edge cases are challenging even with the implementation of the textual context. The misclassification rises according to the random noise that gets introduced by the preprocessing in certain rare circumstances.

Overall, the matrix represents a strong argument for the ETE-ViT's consistent generalization indicating the model captures strong features during training from the Air quality index dataset. This validates the framework's stability and context accurate prediction within the use case of environmental monitoring.

Given the objective of deploying the framework in the real-world settings it is essential to perform an experimentation simulating the greatly unpredictable data fluctuation and noise. Therefore, as shown in Fig. 12 a Noise Robustness Comparison is performed. The test investigates the response to increasing levels of visual noise. The gaussian noise is applied before the preprocessing stage.

As expected, all models show gradual decline in accuracy when noise is increased. Surprisingly ETE-ViT shows the highest resilience. The model maintains a relatively comprehensible semantic text tokenization even when the noise is heavy. This can be justified by the fact that most of the noisy foreground is filtered by the RoI filter and that the sky gets dehazed to a cleaner version. In the extreme noise cases, ETE-ViT retains accuracy above 75%, whereas other models fall below 65% [19], [20].

It can be argued that the integration with VLM text embedding provides a crutch to the transform prediction when pixels are corrupted. The ETE-ViT relies less on pixel fidelity and more on the contextual meaning which results in a robust classifier.

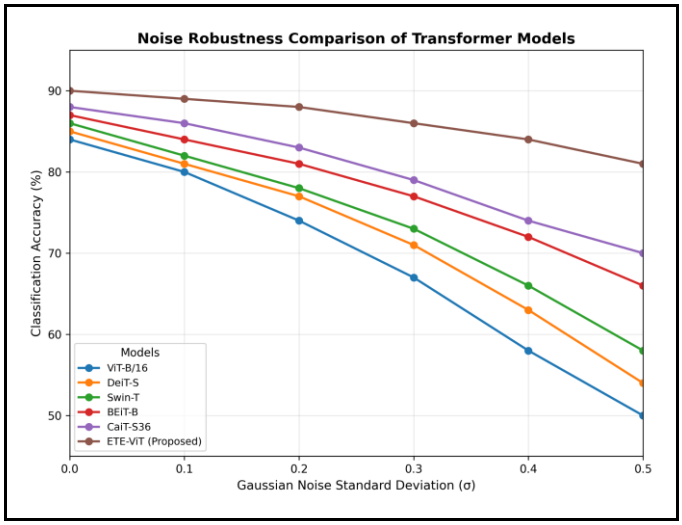


Fig. 12. Noise Robustness Comparison of Transformer Models

Fig. 13 shows the distribution of inference time per module of the ETE-ViT pipeline which provides an idea of the computational footprint of the suggested framework. The Vision Transformer backbone exhibits a fairly efficient

component despite the dense multi head self-attention operations being conducted on high dimensional patch embeddings causing an expected computational cost overhead. Notably, the added multimodal components add a relatively acceptable overhead compared to the baseline ViT. Although the preprocessing stage includes several enhancement algorithms it has a low latency. This is due to leveraging Single Instruction, Multiple Data (SIMD) cpu optimization. Essentially, the Region of Interest filtering enhances the signal to noise ratio of the input data with minor inference time dedication.

The VLM inference has a significant level of computational cost, mainly due to the semantic generation process, but is still within the real time requirement limits. The Guidance Module is the heaviest process requiring 40% of the inference time. The LLM is however asynchronously called and the inference can proceed with further frames while the reasoning is being generated. On the scale of a multi frame inference process this would increase the guidance module time efficiency by up to 50%.

Table III shows the comparison of computational complexity of different transformer architectures tested and the tradeoff between model capacity and accuracy. As can be seen, ETE-ViT does not change the order of magnitude of FLOPs compared to the baseline ViT-B, with around 21 GFLOPs and a moderate parameter increase necessary to enable multimodal integration. This is based on the design decision of maintaining the backbone architecture with the addition of lightweight text embedding mechanisms. The proposed framework is more expensive to use than DeiT-S and Swin-T, but the additional cost is still reasonable. It shows a relevant increase in accuracy.

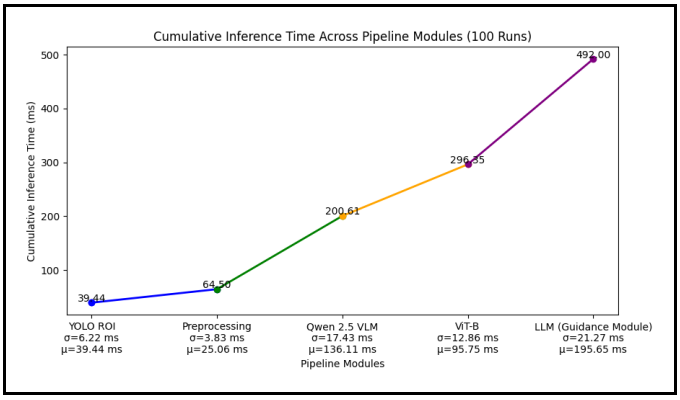


Fig. 13. Inference time per module

TABLE III. MODEL FLOPS COMPARISON

Model	Params (M)	FLOPs (G)	Accuracy
DeiT-S	22	4.6	79.4
Swin-T	29	4.5	80.9
ViT-B	86	17.6	82.1
ETE-ViT	94	21.0	84

Table IV demonstrates an ablation study of the guidance module, which has an incremental effect on the overall performance of the ETE-ViT framework. Each module is sequentially combined with the baseline to demonstrate the accuracy improvement relevant to each component of the pipeline. It is shown the correlation between the preprocessing and the quality of input data which is required in the normalization of feature extraction. The inclusion of text tokens produced by VLM proves the importance of semantic contextualization to direct the attention mechanism on pollution relevant features. This highlights the generalization provided by the semantic tokens. Furthermore, the guidance module improves the relevance of information in the contextual tokens which significantly improve the quality and relevance of the semantic embeddings. Ultimately, the findings support the importance of guided multimodal interaction as a significant variable in enhancing robustness and generalization in the proposed model.

TABLE IV. ABLATION GUIDANCE MODULE

Model	Accuracy
ViT-B baseline	79.8
+ Preprocessing	82.3
+ VLM tokens	83.1
+ Guidance module	84.0

Fig. 14 presents qualitative output samples from the ETE-ViT model. This provides an overview of the classification output across different levels of air pollution before the final semantic module. The model outputs align with ground truth and even succeeds in ambiguous cases where they are even challenging for human perception. It can be concluded that the model displays the required fidelity.

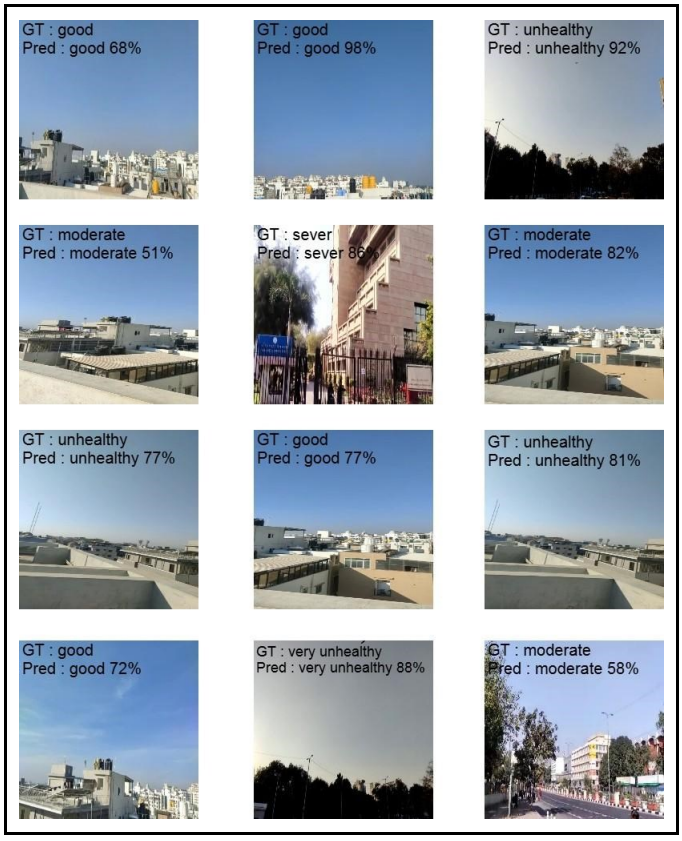


Fig. 14. ETE-ViT Classification Output Samples

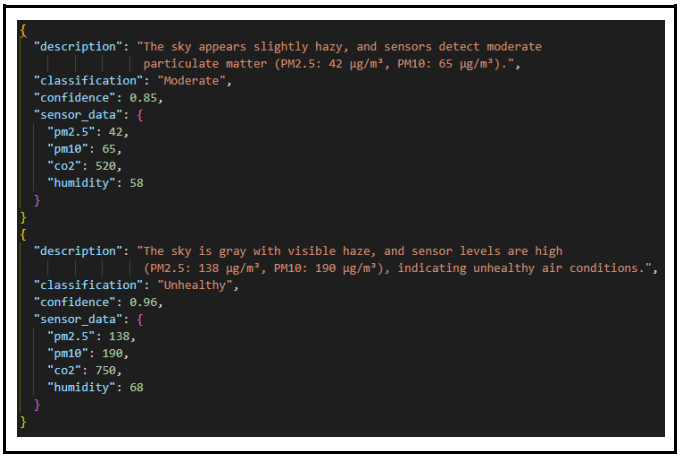
Finally, Fig. 15 demonstrates samples of the final output generated by combining the VLM description and ViT classification. The DeepSeek R1 model associates the results in the format of human interpretable predictions. The contextual statements are concise and clear offering insight on the prediction reasoning.

This feature is essentially Explainable AI providing a glimpse within the chain of thought that led to that specific output. The samples highlight the model's reasoning linking both hardware data with model classification to give a comprehensible output suited for adoption by the end user.

Notably, the improved transparency is a great asset compared to the baseline transformers. The ETE-ViT justifies its prediction which improves the trustworthiness of the classification. Even in cases with high risk of error interpretability, the ability of the model to concisely warn the user about the low confidence or the ambiguity of the prediction becomes a feature instead of a critical failure.

As a final analysis, the experimental results validate the relevance of ETE-ViT architecture to Air pollution classification. The framework demonstrates competitive and superior metrics for this use case. It not only presents a balanced model behavior indicating a successful generalized training but also displays a strong feature extraction. The consistency of the classification across the dataset is validation of reliability.

The ablation analysis does confirm the role of the preprocessing components. This module of the pipeline plays a major role in surpassing the noisy data challenges.



```
{
  "description": "The sky appears slightly hazy, and sensors detect moderate
                  particulate matter (PM2.5: 42 µg/m³, PM10: 65 µg/m³).",
  "classification": "Moderate",
  "confidence": 0.85,
  "sensor_data": {
    "pm2.5": 42,
    "pm10": 65,
    "co2": 520,
    "humidity": 58
  }
}
{
  "description": "The sky is gray with visible haze, and sensor levels are high
                  (PM2.5: 138 µg/m³, PM10: 190 µg/m³), indicating unhealthy air conditions.",
  "classification": "Unhealthy",
  "confidence": 0.96,
  "sensor_data": {
    "pm2.5": 138,
    "pm10": 190,
    "co2": 750,
    "humidity": 68
  }
}
```


Fig. 15. Final Semantic Output Samples

Definitively, the integration of segmenting reasoning layers has leveled the model into the realm of XAI giving it the ability to produce comprehensible outputs instead of the standard classification and confidence score.

Fundamentally, the significance of this research is highlighting the synergy between different models. LLMs, VLMs and ViT drastically benefit from an integration where they support each other and open realms that are usually exclusive to themselves.

The future potential of the framework lies in the scalability and the modular section of the pipeline. A future version that can adapt to new datasets with the intervention of a trainer would represent the peak of accessibility. The potential of cross-domain generalization exists to represent a promising direction for future research. Moreover, the current pipeline has major room for computational optimization which would allow the technology to become valuable in critical infrastructures where milliseconds are critical.

On another note, despite promising results, future exploration is needed in reasoning alignment and enriched multimodal transformers that can handle larger semantic tokens without collapsing the spatial information or the pixel correlation. Additionally, there exists a gap in standard qualitative benchmarking protocol for LLMs and VLMs which can be an interesting research scope.

Overall, the discussion validates that ETE-ViT performance shows both interpretability, and robustness, marking a step forward in multimodal transformer-based perception.

Ultimately, the framework introduces a multimodal pipeline addressing the semantic interpretability limitation identified in the introduction. By converging vision language models encoding with a reasoning mechanism the system dynamically refines the context through prompt augmentation. The results in Fig. 10 and 12 validated the acquired robustness across noisy data and an edge case representation. The resulting improvement in accuracy, stability and interpretability resolve the air pollution challenges raised at the start of the study. The enriched transformer with text tokens features clarity and generalization that displays a balanced distribution across all the classes as shown in Fig. 11. As a closing sentence to this analysis, the results highlight that ETE-ViT not only achieves superior accuracy and interpretability but also establishes a scalable foundation for adaptive multimodal transformers in real world settings.

IV. CONCLUSION

The study introduces ETE-ViT (Enhanced Text-Embedded Vision Transformer) a framework that integrates vision transformers with VLM text embedding with the aim of air pollution classification. The objective is to use semantic textual tokens to improve the flexibility under diverse environmental conditions.

The proposed framework achieves an accuracy of 84% and macro-F1 of 0.823 on the Air Quality index dataset which are competitive across the baseline transformers and also satisfy the requirement of the use case. Between the improved noise robustness and stability across the domain the model displays a strong generalization. The semantic alignment proves the learning improvement associated with the integration of semantic textual tokens.

The results confirm that multimodal transforms enhance the model's feature extraction quality. The contextual anchors serve as a safe guard against noise and uncontrolled environmental factors.

ETE-ViT produces concise semantic predictions that are essential for a real-world adoption of the technology. The interoperability of the framework output lowers the resistance of deployment and drastically improves the accessibility.

To emphasize, the integration of VLMs into ViTs frameworks validates the potential of multimodal pipelines which leverage the synergy of different architectures to broaden the scope of features available.

Notably, the gap in the established standard qualitative benchmarking framework for LLMs leaves room for future research endeavors. Additionally, the question of cross-domain generalization and few shot frameworks are open subject of investigation with valuable contribution to model adaptation and deployment within real world use cases.

ACKNOWLEDGMENT

We would like to thank the authority of Universiti Teknologi Malaysia (UTM) for providing me access to their facilities for research purposes. Notably, we fully acknowledge UTM Matching Grant 04M64 and International Grant (Qiannan Normal University For Nationalities) 1U009 for their financial support that has made this research endeavor possible.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper.

REFERENCES

[1] Pozzer, A., Anenberg, S. C., Dey, S., Haines, A., Lelieveld, J., & Chowdhury, S. (2023). Mortality attributable to ambient air pollution: A review of global estimates. *GeoHealth*, *7*, e2022GH000711. https://doi.org/10.1029/2022GH000711.

[2] Xu, R., Ye, T., Yue, X., Yang, Z., Yu, W., Zhang, Y., Bell, M. L., Morawska, L., Yu, P., Zhang, Y., Wu, Y., Liu, Y., Johnston, F., Lei, Y., Abramson, M. J., Guo, Y., & Li, S. (2023). Global population exposure to landscape fire air pollution from 2000 to 2019. *Nature*, *621*, 521–529. https://doi.org/10.1038/s41586-023-06398-6.

[3] Burke, M. (2021). The changing risk and burden of wildfire in the United States. *Proceedings of the National Academy of Sciences*, *118*(49), e2011048118. https://doi.org/10.1073/pnas.2011048118.

[4] Alfano, B., Spinelle, L., Gerboles, M., & Cattaneo, A. (2020). A review of low-cost particulate matter sensors for air quality monitoring. *Atmosphere*, *11*(2), 212. https://doi.org/10.3390/atmos11020212.

[5] Wang, Y., Li, X., & Zhang, Z. (2023). Surveillance-image-based outdoor air quality monitoring using deep learning. *IEEE Transactions on Geoscience and Remote Sensing*, *61*, 1–12.

[6] Utomo, A. D., *et al.* (2023). Eff-AQI: An efficient CNN-based model for air pollution estimation — A study case in India. In *ACM GoodIT Conference Proceedings*.

[7] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., … & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*. arXiv:2010.11929.

[8] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. *arXiv*. arXiv:2012.12877.

[9] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., … & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv*. arXiv:2103.14030.

[10] Touvron, H., *et al.* (2021). Going deeper with image transformers (CaiT). *arXiv*. arXiv:2106.04560.

[11] Bao, H., Dong, L., & Wei, F. (2021). BEiT: BERT pre-training of image transformers. *arXiv*. arXiv:2106.08254.

[12] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., … & Sutskever, I. (2021). Learning transferable visual models from natural language supervision (CLIP). *arXiv*. arXiv:2103.00020.

[13] Li, J., Li, K., H. H., & Chang, S. (2022). BLIP: Bootstrapping Language-Image Pre-training. *arXiv*. arXiv:2201.12086.

[14] Li, J., et al. (2023). BLIP-2: Bootstrapping Language-Image Pre-training with frozen image encoders and large language models. *arXiv*. arXiv:2301.12597.

[15] Yu, L., *et al.* (2022). CoCa: Contrastive captioners unify vision-language pretraining. *arXiv*. arXiv:2205.01917.

[16] Wang, W., *et al.* (2021). SimVLM: Simple visual language model pre-training with weak supervision. *arXiv*. arXiv:2108.10904.

[17] Li, X., *et al.* (2021). ALBEF: Align before fuse for vision-and-language pretraining. In *Advances in Neural Information Processing Systems (NeurIPS 2021)*.

[18] Alayrac, J.-B., *et al.* (2022). Flamingo: A visual language model for few-shot learning. *arXiv*. arXiv:2204.14198.

[19] Qin, X., *et al.* (2020). FFA-Net: Feature fusion attention network for single image dehazing. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[20] Zamir, S. W., *et al.* (2022). Restormer: Efficient transformer for high-resolution image restoration. *arXiv*. arXiv:2111.09881.

[21] Zhao, Y., *et al.* (2025). YOLOv12: Attention-centric real-time object detectors. *arXiv*. arXiv:2502.12524.

[22] Wei, J., Wang, X., Schuurmans, D., Le, Q., & Makri, A. (2022). Chain-of-thought prompting elicits reasoning in large language models. *arXiv*. arXiv:2201.11903.

[23] Guo, D., *et al*. (2025). DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv*. arXiv:2501.12948.

[24] Maxwell, I. A. (2025). *Meta-Cognitive Prompting: A Comparative Framework for Prompt Engineering in Large Language Models*. ResearchGate preprint. https://doi.org/10.13140/RG.2.2.22405.46562.