Proceedings of the 21st Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 615–620
DOI: 10.15439/2026F1013
ISSN 2300-5963 ACSIS, Vol. 47

Improving Multimodal Plant Disease Captioning through Adaptive Query Token Selection

Rabson Zimba, Fakhia Hammad, Muhammad Hammad Saleem, Mohamad Saraee
0000-0003-4730-0118, 0009-0005-9465-7453, 0000-0002-3625-3021, 0000-0002-3283-1912
School of Science, Engineering, and Environment, University of Salford,
Salford M5 4WT, United Kingdom;
Data Science and Artificial Intelligence Hub,
University of Salford, Salford M5 4WT, United Kingdom
Email: R.Zimba@edu.salford.ac.uk, fakhia.hammad@gmail.com,
{F.Hammad, M.H.Saleem, M.Saraee}@salford.ac.uk

***Abstract*—Recent research has focused on vision-language models (VLMs) that demonstrated high performance, offering significant improvement across a broad range of applications, including agriculture in the era of rapid climate change. However, an important yet underexplored component of VLMs is the role of the query tokens in cross-modal alignment. Most existing VLMs adopt fixed query token configuration, and it is unclear whether this is optimal for crop disease diagnostic tasks. This study systematically evaluates the role of query tokens in VLM-based crop disease captioning, employing a systematic analysis to identify trade-offs and optimal configurations across semantic alignment, lexical accuracy and computational efficiency. Experiments were conducted on the LeafNet dataset using a multimodal framework integrating SigLIP and Qwen2.5 VL. The findings provide clear evidence that both captioning performance and computational cost vary considerably with query-token capacity, showing that no single fixed setup offers the best balance across all objectives. This study highlights the value of adaptive design in multimodal learning and position query token optimisation as a promising direction for future research.**

***Index Terms*—Multimodal learning, vision-language models, plant disease detection, Query token optimisation, image captioning.**

I. Introduction

ACCURATE detection of crop diseases is essential for sustainable agricultural production, as it limits their spread and impact in the field. Deep learning has demonstrated significant advancement in automating disease detection in crops [1]. Recent research has focused on vision-language models (VLMs) that have effectively utilised the integration of images and text by leveraging both visual perception and natural language understanding. Consequently, advances in large-scale multimodal (vision and language) models have facilitated the development of model architectures that incorporate either contrastive learning, generative learning, or both. Contrastive learning models like Contrastive Language-Image Pre-training models (CLIP) enable the joint training of the image encoder (a scaled CNN) and text encoder through a contrastive pre-training task to create a foundation model that has demonstrated success on many datasets [2]. Furthermore, generative learning models, like Flamingo, have enabled the turning of learning into conditional generation or prediction of next tokens given a set of tokens and images [3]. Hybrid models with both contrastive and generative learning, such as BLIP-2, FLAVA, and LLaVA 1.5, have shown high performance in various applications, including agriculture. Their capabilities in image captioning, visual question answering, and semantic interpretation are particularly useful for plant disease diagnosis. Most VLMs use static design assumptions from general benchmarks, particularly in the number and structure of query tokens for aligning vision encoders and Large Language Models. For example, BLIP-2 employs pre-trained frozen image and text encoders connected via a query transformer (Q-former) with a fixed configuration of 32 query tokens to bridge visual and textual representations. While the design has been effective across general applications, it remains unclear whether such default configurations are optimal for domain-specific reasoning tasks. Moreover, research has shown that a fixed configuration in such an architectural setting struggles with fine-grained semantics and capturing task-relevant features effectively [4].

Crop disease diagnosis challenges multimodal reasoning systems due to subtle visual symptoms and variations in lesions, discolouration, and texture. Textual information, including symptom descriptions and agronomic observations, offers valuable cues, but research shows significant variation in VLM's ability to interpret plant disease symptoms. [5]. This shows that these state-of-the-art models may not be fully adapted to domain-specific tasks and requirements. Current advances in VLMs have focused on scaling models [4], tuning instruction [6], prompt engineering [7] and applications with larger and multimodal datasets [8]. However, these methods focus on textual prompts and do not address the visual-language interface, which is essential for conveying visual information before language reasoning. Thus, the number of tokens determines the capacity for cross-modal information; too few tokens may prevent the model from capturing fine-grained visual details of the plant disease, whereas too many tokens may introduce redundancy and increase memory demands.

Motivated by these observations, this study proposes a token ablation analysis on a multimodal framework that integrates a frozen vision encoder, a querying transformer and a large language model for enhanced plant disease diagnosis.

Specifically, the study systematically evaluates the role of query tokens as a cross-modal information bottleneck in multimodal understanding of plant diseases, analysing how the number of queries influences models' performance across semantic alignment, lexical accuracy and computational efficiency. The main contribution of this study is as follows:

1. A systematic assessment of the effects of varying query token configurations on multimodal plant disease diagnostic performance.
2. Analysis of the trade-off characterisation and Pareto-optimal regions between multimodal plant disease captioning performance and computational efficiency.

A. Related Work

VLMs can achieve considerably higher classification accuracy with improvements ranging from 3% to 48.9% over their unimodal counterparts [9][10]. The success and promise of the VLM have driven a paradigm shift from unimodal architectures, thereby spurring similar work on the adoption and optimisation of these large-scale models. Studies on token adaptation in multimodal models can be categorised into transformation-based, similarity-based, attention-based, and query-based methods [11][12]. Transformation-based methods compact visual tokens, similarity-based methods merge clustered tokens, attention-based methods prune tokens based on importance, and query-based methods retain task-relevant information using learned queries [11]. Thus, for plant disease diagnosis, a query-token-based method is suitable as it uses task-relevant language guidance to preserve relevant diagnostic visual evidence from a vision encoder, such as colour, lesions and rust, other than relying on generic visual redundancy and saliency.

Recent work highlights that effective token reduction should both remove irrelevant tokens and preserve those pertinent to the user's query. For example, FlashVLM computes cross-modal similarity between image and text embeddings [13], while methods like Progressive Consistency Distillation (EPIC) and MoVE-KD utilise token consistency and attention-based mechanisms to maintain useful tokens [13][14]. Align-TI further emphasises token multimodal alignment [15]. While FlashVLM focuses on selecting a subset of existing visual tokens at inference time, it does not examine query configurations and allocation during training and validation. Additionally, the token distillation methods applied in MoVE-KD and Align-TI support the argument that efficient VLMs must preserve not only compact representations but also meaningful cross-model token interactions; moreover, these token distillation approaches are limited in agriculture, including plant crop diagnosis.

II. Method

A. Query Token Selection Multimodal Framework

The proposed multimodal query token selection framework combines a pre-trained vision encoder, a querying transformer, and a large language model (LLM) for improved plant disease diagnosis. It features an adaptive token-selection mechanism that maps visual features from images to textual output using an LLM-based Q-former. This approach addresses the static information bottleneck found in traditional multimodal pipelines, creating a dynamic framework for generating accurate captions from input images. Key components include visual feature extraction, query-based cross-modal representation learning, and caption generation, as shown in Fig. 1.

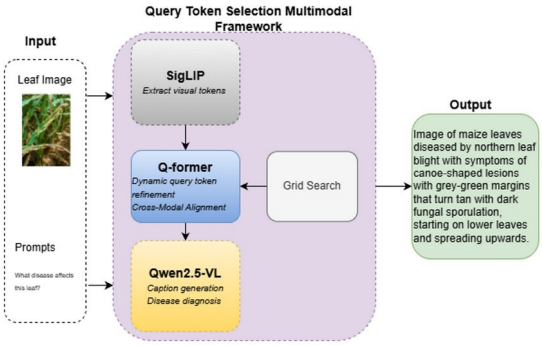


Fig 1. Multimodal framework

B. Visual feature extraction using SigLIP

The proposed framework employed Sigmoid Loss for language-image pretraining (SigLIP)-based vision encoder to obtain high-quality visual representations. SigLIP can learn robust image-text alignment through a sigmoid-based contrastive objective, which is effective for complex domains [16][17]. Given an input image, a SigLIP-based vision encoder extracts representations from the image input, from which it generates a set of visual embeddings representing various regions of the image, such as texture, lesion structures, colour, and patterns. Mathematically, the vision encoder extracts a set of vision tokens (v_i) of a quantity (m) and dimension (d) represented as in Equation 1 below:

$$V = \{v_1, v_2, \ldots, v_m\}, V \in \mathbb{R}^{m \times d} \quad (1)$$

C. Query-based Cross-modal Representation Learning

The Q-former module was adopted to act as a bridge between the vision encoder and the large language model. Q-former has a set of learnable query tokens that attend visual embeddings via cross attention.

$$Q = \{q_1, q_2, \ldots, q_n\} \quad (2)$$

Where Q denotes learnable query tokens, n is the number of query tokens and $n \ll m$.

These learnable query tokens (Q) are employed to interact with visual embeddings to generate multimodal representations (H). These multimodal representations (H) are compressed into a smaller set (n): $H = Qformer(Q, V)$, where $H \in \mathbb{R}^{n \times d}$ represent the cross-modal embeddings. The cross-attention mechanism ($Qformer(Q, V)$ is defined:

$$Attention(QV) = softmax\left(\frac{QW_q(VW_k)^T}{\sqrt{d}}\right)VW_v \quad (3)$$

Where W_q, W_k, W_v are learnable projection matrices, which enable the Q-former to focus on diagnostically relevant visual

regions, compress high-dimensional visual information from the encoder into a smaller number of tokens and facilitate downstream processing by the LLM.

Based on the Information Bottleneck (IB) theory, the multimodal representation (H) should be optimally representative of vision embeddings (V) extracted from Vision Encoder, despite its dimensionality reduction [18]. Thus, the multimodal representation (H) should preserve relevant visual information and should avoid noisy and redundant features. This can be expressed using mutual information.The number of query tokens (n) determines the information capacity of the bottleneck, since the dimensionality of the representations (H) is detected by the number of query tokens. Nevertheless, increasing query tokens (n) increases the information capacity of the bottleneck, allowing more visual information to be conveyed to the LLM. Alternatively, reducing the number of query tokens (n) compressed visual information even more, allowing very little visual information to be conveyed to the LLM. The quality of the output generated by the LLM is influenced by the information (H) conveyed by the Q-former. On one hand, if the capacity of the bottleneck is too small, H may lose useful cues such as lesion structure, spatial patterns or colour variations, which are essential for effective disease diagnosis. On the other hand, too large H increase the demand for computational resources and may introduce redundant features.

D. Query token capacity space

In the conventional fixed token setting, the number of tokens used by the Q-former (K) is fixed; conversely, the study will vary this number by setting it as a search variable from a set of candidates (K).

K = (1, 2, 4, 6, 8, 10, 12, 16, 20, 24, 28, 32, 36, 40, 44, 48, 52, 56, 60, 64, 72, 80, 88, 96, 112, 128).

The token range is selected, as it covers five theoretical regions; severe bottlenecks or low-capacity regions (K = 1 to 12), the optimal region (K = 16 to 64), and the saturated region (K = 72 to 128). The study employed grid search strategies to explore high-performing and yet computationally efficient candidate selection. Thus, grid search was used as a primary reference method for a transparent and complete evaluation of query token candidate values.

E. Caption Generation Using Qwen2.5 VL

The study employed Qwen2.5-VL, a large-scale vision language model capable of handling multimodal inputs and generating contextually rich descriptions. Given the projected multimodal representations (Z), Qwen2.5-VL generates the output sequence autoregressively. The textual prompt inputs are received by the tokeniser of the language model and converted into a set of text tokens (t_i) of a number (k) determined by the prompt length.

$$T = \{t_1, t_2, \ldots, t_k\}, T \in \mathbb{R}^{k \times d} \quad (4)$$

On the other hand, the optimised cross-modal embeddings (H) are projected into the language model's input space. Thus, the language model receives both cross-modal embeddings from Q-former and textual tokens from the prompt.

III. Experiment

To evaluate the effectiveness of the proposed query token framework, a series of experiments was conducted with a focus on two key aspects: the impact of query tokens' bottleneck size on multimodal reasoning and the selection of optimal query tokens using grid search methods. The approach integrates an SIgLIP vision encoder, a Q-former-based query bottleneck and a Qwen2.5-VL language decoder for the generation of captions as outputs. These experiments were conducted on the LeafNet dataset under three captioning settings. The query token selection framework was benchmarked against multiple fixed tokens configurations, and performance was evaluated using standard captioning metrics (BLEU, ROUGE and BERTScore) alongside computational efficacy measures, latency and memory usage.

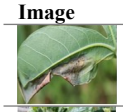
A. Dataset

The study used LeafNet comprehensive multimodal datasets, published in 2026. This dataset contains 121, 337 pairs of RGB plant images and associated descriptions of symptoms and crop type, with up to 97 disease classes [5]. Thus, this multimodal dataset enabled a comprehensive analysis of the performance of an optimised framework on the tasks that require plant disease reasoning. The study employed a cross-validation protocol to reduce dependence on a single train-test. In each fold, one proportion of the dataset was held out as the test set, while the remaining data formed the training-validation pool. Thus, a five-fold cross tabulation was used; each test fold corresponds to 20% of the dataset, and the remaining 80% was used for training, from which 10% was further assigned to validation. The original paired descriptions in the dataset were used as the baseline textual representation, as they exhibit redundancy and inconsistent phrasing. To reduce noise and enhance the alignment between text and images, the captions were cleaned by removing redundant phrases, standardising disease terminology and simplifying the sentence structure. Furthermore, the captions were augmented by adding synonyms and paraphrasing to introduce textual expressions while preserving the underlying semantic meaning, as shown in Table I. The study utilized three caption types (original, cleaned, and augmented) to evaluate multimodal reasoning across various linguistic styles. This approach allowed the model to be tested for robustness against noisy textual descriptions that reflect real-world conditions. The cleaned captions offered a simplified linguistic representation for improved visual-language alignment, while augmented descriptions introduced diversity for better generalisation. All images were resized to 224 x 224 pixels for compatibility with the SigLIP vision encoder, which also handled preprocessing to standardize image input. Additionally, pixel intensity values were normalized to meet the requirements of the transformer-based vision encoder.

B. Training Procedure

To ensure utilisation of limited computer resources, the training was performed using a parameter-efficient setup in which the SigLIP encoder and Qwen language model were

TABLE I. DATA PREPROCESSING WITH ORIGINAL CAPTION, CLEANED CAPTION, AND AUGMENTED CAPTION

Image	Original caption	Cleaned caption	Augmented caption
	An image of Cassava leaves diseased by bacterial blight with symptoms of blight, wilting, dieback, and vascular necrosis	cassava bacterial blight wilting dieback vascular necrosis	the image is consistent with cassava bacterial blight wilting dieback vascular necrosis
	An image of Coffee leaves diseased by Miner with symptoms of Transparent areas in the leaf and larvae maybe present on the underside of the coffee leaf	coffee miner transparent areas larvae underside coffee	observed features are compatible with coffee miner transparent areas larvae underside coffee

frozen, while only the query tokens, cross-attention block and projection layers were optimised. The trainable components of the framework were trained using a sequence-to-sequence objective, minimising the negative log-likelihood to reduce computational requirements while maintaining strong generalisation capabilities. Thus, the training objective was defined as a standard autoregressive cross-entropy loss over the target caption sequence:

$$L = -\sum_{t=1}^{T} \log P\left(y_t \mid y_{<t}, Z\right) \quad (5)$$

where y_t denotes the target token at the time step t, $y_{<t}$ are the previously generated tokens, and Z is the projected multimodal prefix derived from the interaction between the query tokens and the visual features. During the training, the language model generates captions in an autoregressive manner conditioned on both the textual prompt and the visual prefix Z. To stabilise training and accelerate convergence, teacher forcing is employed, where the ground-truth tokens are provided at each time step. To study the influence of each token query configuration set, the model was trained independently with each token value. For every run, the framework recorded the training loss, validation loss, captioning metrics, latency and memory usage. The results from this setup provided direct evidence of how the number of query tokens affects multimodal reasoning and provided a search landscape for the grid search method by defining candidate regions in which the best trade-offs were likely to occur.

C. Evaluation Metrics

To evaluate the performance of the proposed multimodal framework, the study evaluated the semantic quality of generated captions and computational efficiency. Unlike the traditional disease classification tasks, the proposed approach treats plant disease understanding as a multimodal caption generation problem, where the framework must output semantically rich textual descriptions derived from the visual evidence. The study, therefore, employed a set of widely adopted natural language generation metrics, including Bilingual Evaluation Understudy (BLEU), Recall-Oriented Understudy for Gisting Evaluation (ROUGE) and BERTScore. The BLEU metric was used to measure n-gram precision between generated captions and reference texts. Thus, BLEU reflects the degree of lexical overlap, where lower-order n-grams (e.g. BLEU-1) capture word-level correctness, and the higher-order n-grams (e.g. BLEU-4) assess fluency and phrase-level consistency [19][20]. The study employed BLEU to provide an indication of how closely the generated captions match the reference descriptions in terms of surface form.

D. Comparative analysis

To strengthen the external validity, selected query-token configurations were compared against BLIP-2. This included default query tokens in BLIP-2 (32) and an adjusted number of query tokens in the BLIP-2 configuration. The comparison was conducted using the same dataset, prompt, evaluation metrics and held-out test protocol.

E. Environmental Setup

The framework was implemented in Python using PyTorch. The Hugging Face Transformers were employed for the multimodal components. Model evaluation used the Evaluate library, which comprises BLEU, ROUGE and BERTScore. Other libraries used include: TorchVision, PIL, NumPy, Scikit-Learn, Matplotlib and Seaborn. The implementation was designed and executed with careful memory management, which motivated the use of lightweight, trainable components, a quantisation-compatible language model and gradient accumulation. The key components employed during the experiment include: SigLIP vision encoder (kept frozen), Qwen2.5-VL language model (kept frozen), Trainable query tokens with cross attention and projection, custom dataloaders and collate functions and experiment loop with episodic saving capability.

IV. RESULTS AND DISCUSSION

A. Query Variation and Multimodal Learning Behaviour

The experiments show dynamic learning behavior across various token sizes. The training and validation loss reveals that both losses peak at 0.80 with a very small token size (K = 1), indicating limited capacity. As token numbers increase, training loss decreases and stabilizes from K = 48, dropping from 0.75 at K = 1 to lower values starting at K = 20, with reduced fluctuations from K = 48 onward. Validation loss remains low across all sizes, suggesting strong generalization and no overfitting. The fluctuations in loss indicate that query-token capacity affects learning stability and optimization non-monotonically.

B. Multimodal captioning performance vs query token variation

An evaluation of query token sizes using BLEU, ROUGE, and BERTScore metrics assessed crop and disease mention rates. Results indicate that captioning performance is sensitive to query token size. The smallest setting (K = 1–20)

yielded weak scores, while significant improvements were observed at mid-range settings.

The highest BERTScore (0.78), BLEU (0.71), and ROUGE-2 (0.68) occurred at K = 32, balancing semantic and phrase-level overlap. Conversely, ROUGE-1 and ROUGE-L were strongest at K = 56, indicating better lexical overlap and longer correspondences, as presented in Fig. 2. A Wilcoxon signed-rank test revealed significant increases in crop mention rates from 0.04 to 0.23 and disease mentions from 0.47 to 0.68 at K = 56 ($p = 0.047 < 0.05$). Higher token configurations enhance the model's ability to generate accurate agricultural captions, which is vital for effective diagnostic tasks.

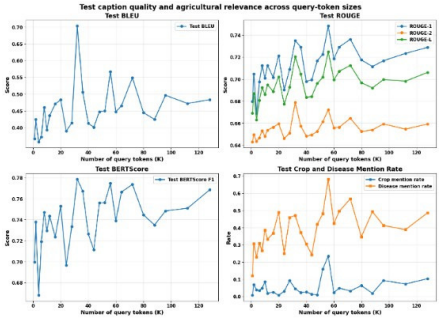


Fig 2. BLEU, ROUGE and BERTScore results

C. Computational efficiency and model's performance Analysis

The experiment results were analyzed to assess the model's efficiency in crop disease diagnosis. As shown in Fig. 3, training time increased with more tokens, rising from 5,712.93 seconds at K=1 to 10,520.94 seconds at K=128. Memory usage also increased significantly, from 6,404.91 MB at K=1 to 9,347.44 MB at K=128. Inference time exhibited no strict pattern, with a low of 0.14 seconds at K=2, peaking at 0.16 seconds at K=60 before decreasing again. Performance analysis revealed a trade-off between semantic accuracy, latency, and memory. K=56 achieved the highest ROUGE-1 score of 0.27 and an efficiency score of 2.48, indicating it offered the best overall performance. However, it required 8,968.69 seconds of training and 8,266.03 MB of GPU memory, while K=32 offered a more balanced option with lower computational costs and 0.15 seconds of latency. The experiment results were analyzed to assess the model's efficiency in crop disease diagnosis. As shown in Fig. 3, training time increased with more tokens, rising from 5,712.93 seconds at K=1 to 10,520.94 seconds at K=128. Memory usage also increased significantly, from 6,404.91 MB at K=1 to 9,347.44 MB at K=128. Inference time exhibited no strict pattern, with a low of 0.14 seconds at K=2, peaking at 0.16 seconds at K=60 before decreasing again. Performance analysis revealed a trade-off between semantic accuracy, latency, and memory. K=56 achieved the highest ROUGE-1 score of 0.27 and an efficiency score of 2.48, indicating it offered the best overall performance. However, it required 8,968.69 seconds of training and 8,266.03 MB of GPU memory, while K=32 offered a more balanced option with lower computational costs and 0.15 seconds of latency.

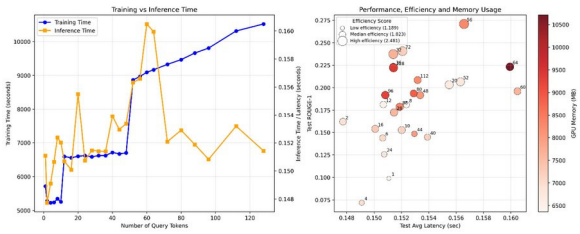


Fig 3. Plots for computational efficiency and performance

D. Comparative analysis

To externally compare the influence of query tokens on performance, efficiency and performance, the study conducted an ablation analysis on the BLIP-2 architecture on selected query tokens. The findings, summarised in Table II, support the central argument that a fixed query token is not ideal for agricultural diagnostic captioning. Thus, Unlike the framework used in the study, BLIP-2 query tokens 16 and 48 provided strong metrics across BLEU, ROUGE and combined rate for crop and disease mention. Thus, the variation in metrics in BLIP further justifies treating query token selection as a multi-objective design problem rather than adopting a fixed design.

TABLE II: BLIP-2 QUERY TOKEN PERFORMANCE.

Query tokens (K)	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	Combined mention	Latency/ image (s)	Training time (h)
4	0.1602	0.3977	0.1962	0.3841	0.0000	**0.4033**	**3.27**
16	0.3377	0.4741	**0.3635**	0.4699	0.1125	0.4141	3.40
32	0.3438	0.4752	0.3632	0.4715	0.1138	0.4303	3.68
48	**0.3499**	**0.4763**	0.3629	**0.4730**	**0.1150**	0.4464	3.96
56	0.3350	0.4697	0.3537	0.4659	0.0863	0.4577	4.06
96	0.2902	0.4499	0.3191	0.4445	0.0000	0.4905	4.21

E. Discussion

The study evaluated whether a fixed token configuration is sufficient for agricultural image caption with focus on plant disease description, diagnostic relevance and computational efficiency. The experiment provides clear evidence that query tokens' dimensionality plays a fundamental role in influencing multimodal reasoning performance. Across all evaluated configurations, performance exhibits a non-linear relationship with query token size. However, the initial phase saw improvement in various metrics, followed by diminishing returns. Specifically, a lower token configuration (K<20) is associated with reduced representational capacity, leading to suboptimal semantic alignment and higher validation loss. On the other hand, increasing the number of query tokens enhances the richness of multimodal representations, enabling more accurate capture of fine-grained disease features in the leaf images. This finding aligns with prior findings in the transformer-based architecture, where increasing representational capacity enhances performance up to a saturation point beyond which redundancy is introduced [21]. Furthermore, the results demonstrate that this improvement might not be definitive. Beyond a certain level, additional tokens yield marginal gains relative to computational cost. The emergence of K=56 as the best configuration when captioning

quality and performance are prioritised indicates the existence of a critical capacity point, especially for complex image features like those of crop disease, where the model achieves reasonable representation efficiency without unnecessary redundancy. Besides, the finding is consistent with the theoretical perspective on information bottlenecks in deep learning systems.

The experiment results demonstrate the evidence that fixed token configuration may be inadequate for agricultural captioning. Thus, different token values favour different objectives, including captioning overlap, semantic similarity, diagnostic relevance, validation suitability and computational efficiency. Moreover, prior research work has shown that hyperparameter sensitivity in multimodal architecture can significantly influence downstream performance, including in cross-modal alignment [2][3][16].

V. Conclusion and Future Work

The study evaluates the impact of query token dimensionality in multimodal vision-language models and systematic analysis on an adaptive query selection framework. The results demonstrate that query token dimensionality significantly influences both performance and efficiency, and that different token configuration settings favour different objectives. This study highlights the importance of adaptive design in multimodal learning and establishes query token optimisation as a promising direction for future research. This work has some limitations. optimisation was conducted over a discrete set of token configurations, which may not fully capture the continuous design space. Second, a selected number of tokens were evaluated across different models, limiting the ability to conduct an extended comparison across the token search space and different architectures such as InstructBLIP and LLaVA. Future work will build on this by incorporating repeated training runs across architectural designs to formulate a robust multimodal framework for enhancing crop disease diagnostics. Further extending this work into a fully explainable multimodal framework, adapted per input sample, as well as joint optimisation of multiple architectural parameters.

Generative AI Disclosure Statement

Grammarly was used to improve language of this manuscript.

References

[1] Saleem, M.H., Potgieter, J., and Arif, K.M.: 'Automation in Agriculture by Machine and Deep Learning Techniques: A Review of Recent Developments', Precision Agriculture, 2021, 22, (6), pp. 2053–2091. DOI: 10.1007/s11119-021-09806-x

[2] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., and Clark, J.: 'Learning transferable visual models from natural language supervision', in: 'Book Learning transferable visual models from natural language supervision' (PmLR, 2021, edn.), pp. 8748–8763.

[3] Awais, M., Naseer, M., Khan, S., Anwer, R.M., Cholakkal, H., Shah, M., Yang, M.H., and Khan, F.S.: 'Foundation Models Defining a New Era in Vision: A Survey and Outlook', IEEE Trans Pattern Anal Mach Intell, 2025, 47, (4), pp. 2245–2264. DOI: 10.1109/TPAMI.2024.3506283

[4] Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., and Hoi, S.: 'Instructblip: Towards general-purpose vision-language models with instruction tuning', Advances in neural information processing systems, 2023, 36, pp. 49250–49267.

[5] Quoc, K.N., Dao, P.D., and Quach, L.-D.: 'LeafNet: A Large-Scale Dataset and Comprehensive Benchmark for Foundational Vision-Language Understanding of Plant Diseases', arXiv preprint arXiv:2602.13662, 2026

[6] Dong, J., Yao, Y., Fuentes, A., Jeong, Y., Yoon, S., and Park, D.S.: 'Visual information guided multi-modal model for plant disease anomaly detection', Smart Agricultural Technology, 2024, 9, pp. 100568. DOI: 10.1016/j.atech.2024.100568

[7] Ghosh, D., Ghosh, S., Jana, N.D., and Mallipeddi, R.: 'Intelligent neural architecture search via Taguchi design and language model-based differential evolution for agricultural image recognition', Expert Systems with Applications, 2026, 309, pp. 131135. DOI: 10.1016/j.eswa.2026.131135

[8] Bie, Y., Xu, G., and Wang, Y.: 'Adapting Vision-Language Models for Precision Agriculture: A Study on Crop Segmentation based on UAV Remote Sensing Data', in 'Book Adapting Vision-Language Models for Precision Agriculture: A Study on Crop Segmentation based on UAV Remote Sensing Data' (2025, edn.), pp. 1–6. DOI: 10.1109/Agro-Geoinformatics66479.2025.11136213

[9] Upadhyay, A., Chandel, N.S., Singh, K.P., Chakraborty, S.K., Nandede, B.M., Kumar, M., Subeesh, A., Upendar, K., Salem, A., and Elbeltagi, A.: 'Deep learning and computer vision in plant disease detection: a comprehensive review of techniques, models, and trends in precision agriculture', Artificial Intelligence Review, 2025, 58, (3), pp. 92. DOI: 10.1007/s10462-024-11100-x

[10] Xu, C., Zhao, L., Wen, H., Zhang, Y., Wang, L., and Zhang, L.: 'Intelligent management of crop diseases and pests in multiscale and multimodal complex scenarios: Technologies, applications, and prospects', Computers and Electronics in Agriculture, 2026, 244, pp. 111443. DOI: 10.1016/j.compag.2026.111443

[11] Shao, K., Keda, T., Zhang, K., Feng, S., Cai, M., Shang, Y., You, H., Qin, C., Sui, Y., and Wang, H.: 'A Survey of Token Compression for Efficient Multimodal Large Language Models', Transactions on Machine Learning Research, 2026

[12] Jin, Y., Li, J., Gu, T., Liu, Y., Zhao, B., Lai, J., Gan, Z., Wang, Y., Wang, C., Tan, X., and Ma, L.: 'Efficient multimodal large language models: a survey', Visual Intelligence, 2025, 3, (1), pp. 27. DOI: 10.1007/s44267-025-00099-6

[13] Cai, K., Zhang, J., Yang, J., Fan, Y., Xie, P., Wang, J., and Wang, K.: 'Flashvlm: Text-guided visual token selection for large multimodal models', arXiv preprint arXiv:2512.20561, 2025

[14] Wen, F., Guo, Y., Wang, J., Xiao, J., Zhou, Y., Shen, Y., Li, C., Jia, Q., and Zhang, Z.: 'Improve MLLM Benchmark Efficiency Through Interview', in 'Book Improve MLLM Benchmark Efficiency Through Interview' (2026, edn.), pp. 19892–19896. DOI: 10.1109/ICASSP55912.2026.11464672

[15] Chen, L., Zhao, X., Ding, K., Feng, W., Miao, C., Wang, Z., Guo, W., Wang, Y., Zheng, K., and Zhang, B.: 'Beyond next-token alignment: Distilling multimodal large language models via token interactions', arXiv preprint arXiv:2602.09483, 2026

[16] K. M, S., Kumar, H., and Charangatt, J.V.: 'Disease Classification and Report Generation from Chest X-Ray Images Using SigLIP and LLaVA', in 'Book Disease Classification and Report Generation from Chest X-Ray Images Using SigLIP and LLaVA' (2025, edn.), pp. 1514–1518. DOI: 10.1109/TENCON66050.2025.11375249

[17] Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L.: 'Sigmoid loss for language image pre-training', in 'Book Sigmoid loss for language image pre-training' (2023, edn.), pp. 11975–11986.

[18] Goldfeld, Z., and Polyanskiy, Y.: 'The Information Bottleneck Problem and its Applications in Machine Learning', IEEE Journal on Selected Areas in Information Theory, 2020, 1, (1), pp. 19–38. DOI: 10.1109/jsait.2020.2991561

[19] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J.: 'Bleu: a method for automatic evaluation of machine translation', in: 'Book Bleu: a method for automatic evaluation of machine translation' (2002, edn.), pp. 311–318.

[20] Hrga, I., and Ivasic-Kos, M.: 'Measuring the Sensitivity of Image Captioning Metrics to Caption Perturbations', in 'Book Measuring the Sensitivity of Image Captioning Metrics to Caption Perturbations' (Springer Nature Singapore, 2024, edn.), pp. 1053–1063.

[21] Chen, J., Wu, P., Zhang, X., Xu, R., and Liang, J.: 'Add-Vit: CNN-Transformer Hybrid Architecture for Small Data Paradigm Processing', Neural Processing Letters, 2024, 56, (3), pp. 198. DOI: 10.1007/s11063-024-11643-8