VLM Fine-tuning and Multi-Model Ensemble for Synthetic Images Detection

Hengwei Zhao[1], Takayuki Hori[1]

[1] *SoftBank Corporation, Tokyo, Japan*

Abstract

This paper describes our participation in the MediaEval 2026 Synthetic Images Task, Subtask A: Synthetic Image Detection. The objective of this task is to classify images as either real or synthetic. We explored two different approaches. For the constrained run, we fine-tuned Vision-Language Models (VLMs), namely Qwen3-VL-8B-Instruct and Qwen3-VL-32B-Instruct [1], using the official training data [4]. For the open run, we investigated a multi-model ensemble framework combining Qwen2.5-VL-32B-Instruct, Qwen3-VL-32B-Instruct, GPT-4o, and Gemma-3-27b-it [2,3]. Experimental results show that VLM fine-tuning can achieve competitive performance, while ensemble-based approaches improve accuracy and generalization on unseen test data. Our analysis suggests that model diversity and decision aggregation are effective strategies for synthetic image detection.

1 Introduction

As image generation models become increasingly realistic, reliable detection methods are required to distinguish synthetic content from authentic photographs.

In this work, we participated in the MediaEval 2026 Synthetic Images Task, Subtask A: Synthetic Image Detection. The goal of this task is to classify an image as either real or synthetic. Detailed descriptions of the task, datasets, and evaluation protocol are provided in the task overview paper [4].

Our participation focused on two directions. First, under the constrained setting, we investigated supervised fine-tuning of large Vision-Language Models (VLMs) using only the provided training data. Second, under the open setting, we explored a multi-model ensemble framework that combines predictions from multiple models. The objective was to evaluate whether ensemble reasoning can improve accuracy when detecting synthetic images.

The main contributions of our work are:

- Evaluation of Qwen3-VL models for synthetic image detection under constrained conditions.
- Investigation of training set size effects on VLM fine-tuning.
- Design of a multi-model ensemble framework for open run.
- Analysis of performance differences between validation and official test results.

2 Related Work

Synthetic image detection has attracted significant attention due to advances in diffusion models and large-scale image generators [5]. Previous approaches can be broadly categorized into frequency-domain analysis, and deep learning classifiers.

Recently, Vision-Language Models have demonstrated strong multimodal understanding capabilities and have been applied to image authenticity assessment [6,7]. In addition, ensemble reasoning has been shown to improve accuracy by combining complementary decision patterns from multiple models [8].

Motivated by these findings, we investigate both VLM fine-tuning and ensemble-based decision aggregation for synthetic image detection.

3 Approach

3.1 Constrained Run: VLM Fine-Tuning

For the constrained run, we fine-tuned two Vision-Language Models with LoRA Fine-tuning [9]:

- Qwen3-VL-8B-Instruct
- Qwen3-VL-32B-Instruct

[1]*MediaEval'26: Multimedia Evaluation Workshop, June 15–16, 2026, Amsterdam, Netherlands and Online*

*Corresponding author.

hengwei.zhao@g.softbank.co.jp (H. Zhao); takayuki.hori@g.softbank.co.jp (T. Hori)

The training configurations are shown below:

ID	Model	Training Images
Cons_run_1	Qwen3-VL-8B-Instruct	4,000
Cons_run_2	Qwen3-VL-8B-Instruct	20,000
Cons_run_3	Qwen3-VL-32B-Instruct	4,000
Cons_run_4	Qwen3-VL-32B-Instruct	20,000

The datasets contained equal numbers of real and synthetic images. Validation datasets were also balanced and consisted of 400 and 2,000 images, respectively.
The task was formulated as binary classification. During inference, the model generated a prediction indicating whether the input image was real or synthetic.

3.2 Open Run: Multi-Model Ensemble

For the open run, we designed a multi-model ensemble framework.
The following models were evaluated:

- Qwen2.5-VL-32B-Instruct
- Qwen3-VL-32B-Instruct
- GPT-4o
- Gemma-3-27b-it

We experimented with two aggregation strategies:

- **Majority Voting**
  Each model produced a binary prediction. The final decision was determined by majority vote.
- **Probability Averaging**
  Each model generated a confidence score. The final prediction score was obtained by averaging confidence values across models. Different decision thresholds were evaluated to maximize validation F1 score.

Several ensemble combinations were explored:

- Qwen2.5 + Qwen3 + GPT-4o
- Qwen3 + GPT-4o
- Qwen2.5 + Qwen3 + GPT-4o + Gemma

4 Results and Analysis

4.1 Constrained Run Results

Table 1: Results of Constrained Run
The threshold is 0.5 for all runs. Performance is evaluated using metrics where Accuracy, Precision, Recall, and F1-score are abbreviated as Acc, P, R, and F1, respectively.

	Validation				Test			
ID	Acc	P	R	F1	Acc	P	R	F1
Cons_run_1	**0.9327**	0.9072	0.9642	**0.9348**	**0.8889**	0.9459	0.8250	0.8813
Cons_run_2	0.6494	0.5887	**0.9920**	0.7389	0.6171	0.9487	0.2476	0.3927
Cons_run_3	0.9008	**0.9491**	0.8472	0.8952	0.8870	0.8703	**0.9096**	**0.8895**
Cons_run_4	0.8522	0.7764	0.9894	0.8701	0.8378	**0.9837**	0.6870	0.8090

Among constrained runs, Qwen3-VL-32B trained on 4,000 images achieved the best official test performance with a F1 score of 0.8895. In contrast, Qwen3-VL-8B trained on 20,000 images showed a substantial performance drop, obtaining an F1 score of 0.3927 on the official test set.
This observation indicates that increasing the training data size does not necessarily improve generalization. The larger model appeared to generalize better with limited but balanced training data.

4.2 Open Run Results

Table 2: Results of Open Run
The best threshold was evaluated to maximize validation F1 score for probability averaging. Performance is evaluated using metrics where Accuracy, Precision, Recall, and F1-score are abbreviated as Acc, P, R, and F1, respectively.

			Validation				Test			
Approach	Model	threshold	Acc	P	R	F1	Acc	P	R	F1
Single model	Qwen2.5	0.5	0.9117	0.9285	0.9256	0.9271	0.8730	0.9046	0.8340	0.8678
Single model	Qwen3	0.5	0.9399	0.9405	0.9618	0.9510	0.8954	0.9435	0.8412	0.8894
Single model	GPT-4o	0.5	0.9455	**0.9842**	0.9250	0.9537	0.9071	0.9082	**0.9058**	**0.9070**
Single model	Gemma	0.5	0.8824	0.9235	0.8790	0.9007	0.8580	0.8688	0.8434	0.8559
Majority Voting	Qwen2.5 + Qwen3 + GPT-4o	0.5	0.9523	0.9606	0.9608	0.9607	0.9088	**0.9438**	0.8694	0.9051
Probability Averaging	Qwen2.5 + Qwen3 + GPT-4o	**0.4**	0.9544	0.9641	0.9606	0.9623	**0.9099**	0.9426	0.8730	**0.9064**
Probability Averaging	Qwen3 + GPT-4o	**0.505**	**0.9590**	0.9688	**0.9634**	**0.9661**	0.9087	0.9409	0.8722	0.9052
Probability Averaging	Qwen2.5 + Qwen3 + GPT-4o + Gemma	**0.5**	0.9419	0.9536	0.9504	0.9520				

Among individual models, GPT-4o achieved the highest official test F1 score of 0.9070. Qwen3 achieved a comparable F1 score of 0.8894, while Gemma obtained 0.8559.
The best validation performance was achieved by the Qwen3 + GPT-4o ensemble with probability averaging, reaching a F1 score of 0.9661. However, on the official test set, the performance decreased to approximately 0.9052.
The ensemble of Qwen2.5 + Qwen3 + GPT-4o with probability averaging, achieved an official test F1 score of 0.9064, which was comparable to GPT-4o alone.

4.3 Analysis

Several insights emerged from our experiments.
First, validation performance was not always predictive of official test performance. Some models that achieved very high validation scores experienced notable degradation on unseen test data.
Second, adding more models to an ensemble did not consistently improve performance. Including Gemma generally reduced validation performance, suggesting that weaker or less calibrated models may introduce noise into ensemble predictions.
Third, GPT-4o demonstrated strong standalone performance and remained highly competitive against more complex ensemble strategies. This result suggests that large multimodal foundation models possess strong intrinsic capabilities for synthetic image detection.
Finally, threshold selection played a critical role in probability averaging ensembles. Small threshold adjustments significantly affected F1 score, highlighting the importance of calibration when deploying ensemble systems.

5 Discussion and Outlook

This paper presented our participation in the MediaEval 2026 Synthetic Images Task. We investigated VLM fine-tuning for constrained run and multi-model ensembles for open run.

Our results demonstrate that Vision-Language Models can effectively perform synthetic image detection and that ensemble reasoning can further improve accuracy. However, the discrepancy between validation and official test results indicates that distribution shifts remain a major challenge.

In future work, we plan to investigate adaptive ensemble weighting, and synthetic image attribution methods. We also aim to explore hybrid approaches that combine multimodal reasoning with image forensic features.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to: grammar, spelling and phrasing check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

REFERENCES

[1] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, et al., Qwen3-VL Technical Report, arXiv preprint arXiv:2511.21631, 2025.

[2] OpenAI, GPT-4o System Card, arXiv preprint arXiv:2410.21276, 2024.

[3] Gemma Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, et al., Gemma 3 Technical Report, arXiv preprint arXiv:2503.19786, 2025.

[4] O. Papadopoulou, D. Karageorgiou, C. Koutlis, E. Gavves, H. Mareen, S. Papadopoulos, Synthetic Images at MediaEval 2026: Advancing Detection of Generative AI in Real-World Online Images, Proceedings of MediaEval'26: Multimedia Evaluation Workshop, Amsterdam, Netherlands and Online, 2026

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-Resolution Image Synthesis with Latent Diffusion Models, Proceedings of CVPR, 2022.

[6] A. Radford, J. W. Kim, C. Hallacy, et al., Learning Transferable Visual Models From Natural Language Supervision, Proceedings of ICML, 2021.

[7] S. Bai, K. Chen, X. Liu, et al., Qwen2.5-VL Technical Report, arXiv preprint arXiv:2502.13923, 2025.

[8] T. G. Dietterich, Ensemble Methods in Machine Learning, Multiple Classifier Systems, Springer-Verlag Berlin Heidelberg, 2000.

[9] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, LoRA: Low-Rank Adaptation of Large Language Models, arXiv preprint arXiv:2106.09685, 2021.