

Team CPS Lab at SHROOM-Visions: Parameter-Efficient Vision-Language Hallucination Detection with Threshold Tuning

Anika Srivastava^{1,*}, Vishal Kumar^{1,†}, Harshita Sharma^{1,‡}, Nitisha Aggarwal^{1,§},
Geetika Jain Saxena^{2,¶}, Amit Pundir^{2,||}, Sanjeev Singh^{1,**}

¹Institute of Informatics and Communication, University of Delhi

²Department of Electronics, Maharaja Agrasen College, University of Delhi

*anikasrivastava151@gmail.com, †researchwithvishal@gmail.com, ‡harshitasharma@south.du.ac.in,
§nitisha@south.du.ac.in, ¶gsaxena@mac.du.ac.in, ||amitpundir@mac.du.ac.in, **sanjeev@south.du.ac.in

Abstract

Fine-grained large-vision-language hallucination detection presents the challenge of locating hallucinated content in image-grounded responses and classifying them with corresponding hallucination categories. Our method for the SHROOM-Visions 2026 Shared Task explores a parameter-efficient solution by fine-tuning a lightweight token-level classification head on the frozen weights of the Qwen3-VL-8B-Instruct vision-language model. We utilized category-specific thresholds calculated based on validation character-level IoU and designed a span merging post-processing step to facilitate span localization with fewer fragmented predictions. We also establish a possible connection between the annotator agreement and the operating thresholds. Our analysis suggests that, within our experimental setup, hallucination categories that had higher inter-annotator agreement tended to have higher optimal decision thresholds whereas categories with lower agreement tended to have lower optimal decision thresholds. Our parameter-efficient method achieved ranks 18th, 16th, 16th, and 15th place on the English, French, Italian, and Chinese tracks, respectively, showing promising performance for multilingual fine-grained hallucination detection.

1 Introduction

Large Vision Language Models (LVLMs) such as BLIP-2 (Li et al., 2023) and LLaVA (Liu et al., 2023) have achieved remarkable capabilities in Visual Question Answering problems. LVLMs jointly model visual and textual information. These models generate fluent and contextually aware responses that have enabled their development in real-world applications. Even with these improvements, LVLMs might create objects that are not actually in the image, give wrong descriptions of content, or misunderstand text that is clearly shown in the image (Rohrbach et al., 2019) (Li et al., 2023)

(Wang et al., 2024b) (Gunjal et al., 2024). These erroneous generations, referred to as hallucinations, undermine the reliability of multimodal systems. Hallucinations are critical in applications where inaccurate visual descriptions may lead to incorrect decisions. This creates a direction of research in developing reliable techniques for detecting hallucinated content generated by the LVLMs.

The SHROOM-visions shared task (Vázquez et al., 2026) addresses this challenge by introducing a model agnostic multilingual benchmark for fine-grained hallucination detection in image-conditioned text generation (Mickus et al., 2026). This shared task is required to identify hallucinated character spans and classify each hallucinated span into one of five categories: Invention, Mischaracterization, Optical Character Recognition (OCR) Problem, Miscounting and Other. Unlike conventional sequence level classification problems, the task requires character-level localization along with calibrated confidence estimation, which makes it more difficult than binary hallucination detection problems. While fine-tuning a vision-language model (VLM) achieves good performance, it is computationally expensive and requires updating a large number of parameters (Prattasha et al., 2025).

In this work, we present a lightweight framework for fine-grained VLM hallucination detection built over a frozen pretrained VLM. We employed Qwen3-VL-8B-Instruct (Bai et al., 2025) as the vision-language encoder for joint image-prompt-response embedding, and designed a token-level classifier head for hallucination span detection. To better localize the hallucination spans, we further developed a threshold optimization strategy specific to each hallucination category and applied span merging post-processing. Our method demonstrated that parameter-efficient adaptation can be effective, with less than 84 million parameters being optimized, which is about 1% of the total parameters in the backbone model.

2 Related Work

LVLMS including BLIP-2 (Li et al., 2023), LLaVA (Liu et al., 2023) and Qwen3-VL (Bai et al., 2025) have improved accuracy in multimodal tasks. Still, they suffer from inconsistency in their generated outputs in hallucinated content. Existing studies have investigated hallucinations and highlighted the need for methods that improve factual consistency between generated text and visual content (Li et al., 2023) (Wang et al., 2024b). Previous hallucination detection approaches engineer this problem as response-level classification.

(Rohrbach et al., 2019) demonstrated that as the performance of image captioning models has steadily increased, they tend to hallucinate more, reporting objects that are not present in the scene. For this reason, object hallucination has become a major concern in the field of vision-language generation. With the advent of LVLMS, this problem has grown more severe as such models start producing long-winded responses based on the input image. AMBER (Wang et al., 2024a) introduced an LLM-free multidimensional benchmark for hallucination detection in LVLMS. MHALO (Cai et al., 2025) focuses on fine-grained hallucination detection in LVLMS. It evaluates the model’s capacity to detect hallucinated tokens at the token level for different categories of hallucination. The results presented by the authors demonstrated that current LVLMS still face challenges in detecting fine-grained hallucinations effectively.

Fine-grained methods have recently emerged as a better alternative for hallucination detection problems. Instead of predicting a single label, span-based methods identify the exact region in the generated output responsible for incorrect information. However, accurately detecting hallucinating spans remains challenging. The SHROOM-vision benchmark (Mickus et al., 2026) advances this direction by introducing multilingual character-level span annotations together with fine-grained hallucination categories, encouraging the development of robust multimodal span detection systems.

3 Methodology

Figure 1 presents the overall workflow of the proposed system. The methodology consisted of five phases, namely, dataset preparation and alignment, multimodal model and classifier construction, training and validation, threshold tuning and span post-processing, and final test-file generation.

3.1 Dataset Preparation and Alignment

The training data from English, French, Italian, and Chinese were combined into one multilingual dataset. The data were split into training and validation sets with an 80:20 ratio at the image level to avoid data leakage from one prompt used with the same image. Therefore, all examples that used the same image were placed in one split.

The span was defined as a character-level continuous span in the generated response that was annotated as containing hallucinations. Each span was categorized into one of the five types of hallucination: Invention, Mischaracterization, Miscalculation, OCR, and Other. Since annotations were provided at the character level, they were converted to token-level spans using character offsets.

The multimodal inputs were converted into the chat format, as expected by the VLM. The images were resized with the aspect ratio preserved to have at most 896×896 pixels. To accelerate training, the processed text samples and image representations were cached. Finally, a custom collator was used to batch the cached samples with padded sequences while preserving the token labels and loss masks.

For each token, a multi-hot soft-label vector was constructed from the hallucination spans overlapping with this token, which allowed a single token to have more than one hallucination label if it was overlapped by multiple spans. Additionally, a loss mask was created to ensure that only response tokens contribute to the loss.

3.2 Multimodal Model and Classifier

Qwen3-VL-8B-Instruct was utilized as the pre-trained vision-language backbone. The model encodes the input image and the text prompt-response into contextualized representations for the response tokens. The entire pretrained backbone is frozen during training.

A lightweight trainable classification head was placed on top of these representations. The head consists of a dropout layer, a one-dimensional convolution to model local span coherence, a GELU activation, and a linear projection to logits over the five hallucination categories, trainable parameters of which are shown in Appendix Table A.1. The classifier predicts independent soft confidence scores for each category per response token.

3.3 Training and Validation

Only the parameters of the classification head were optimized. The classifier was trained with a masked

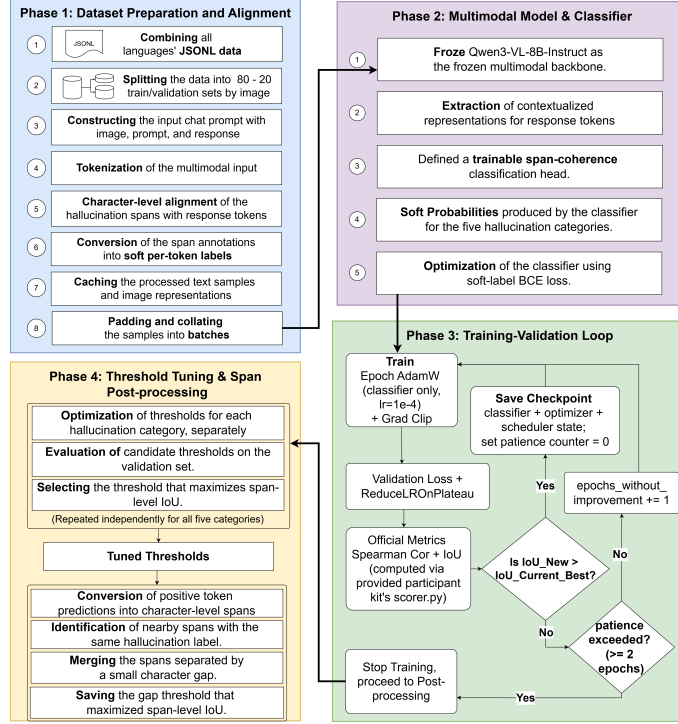


Figure 1: Process Flow of the Proposed Methodology for Hallucination Detection

binary cross-entropy loss, using the soft annotation probabilities as targets. All experiments were performed with the AdamW optimizer, with a learning rate of $1e-4$ and gradient clipping for stability.

After each training epoch, the model was evaluated on the validation set in terms of the official evaluation criteria – Spearman Correlation and IoU. The model checkpoint with the best IoU on the validation set was used for saving. A ReduceLROnPlateau scheduler was used to lower the learning rate if the validation loss stopped improving. It was observed that the training process met the early-stopping condition at the starting learning rate of $1e-4$. Early stopping was triggered based on the validation IoU, with training stopping when the validation IoU did not improve for the specified patience period of 2. The best-performing checkpoint was subsequently used for the post-processing and test-generation stages.

3.4 Threshold Tuning & Span Post-processing

The classifier produced continuous confidence scores for each hallucination category. These scores were converted into binary token predictions using category-specific thresholds estimated on the validation set.

The ground-truth token labels are represented as soft probabilities obtained from annotator agree-

ment. To obtain thresholds, these soft labels were first binarized using a threshold of 0.5. For each hallucination category independently, a precision-recall curve was computed using the validation predictions, and the decision threshold that maximized the token-level F1 score was selected. This procedure resulted in one threshold for each hallucination category.

The resulting binary token predictions were mapped back to character positions using the tokenizer’s offset mappings. Consecutive positive tokens of the same hallucination category were then grouped into a single character-level span. To minimize fragmented predictions, adjacent spans with the same hallucination category but a small character gap between them were merged in a post-processing step.

Finally, the confidence score for each predicted span was calculated as the mean of sigmoid probability of all tokens that belong to the span. Since the classifier is multi-label, for a given token it is possible to have multiple hallucination categories that simultaneously exceed the threshold. In these cases, spans of different categories were kept independent of each other, i.e., without resolving label conflicts. The impact of category-specific threshold tuning and span merging on the shared task’s evaluation metrics is shown in Figure 2.

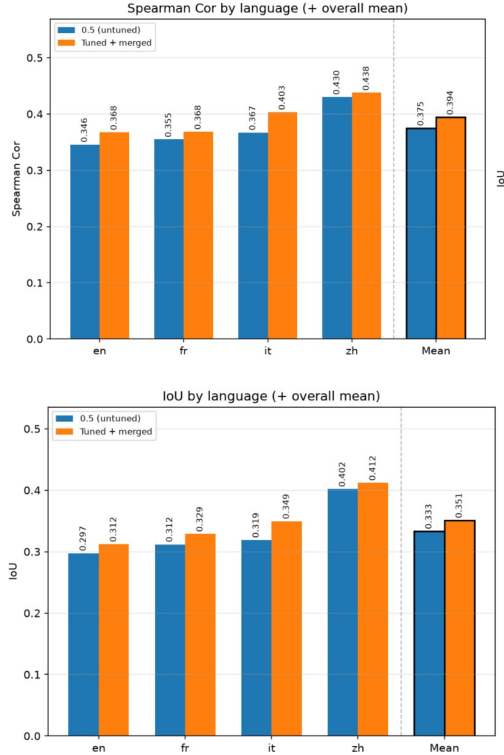


Figure 2: Comparison of model performance before and after threshold tuning and span-merging. Spearman Correlation and IoU show improvement after threshold tuning and span-merging across languages.

3.5 Test File Generation

For the final evaluation, the best classifier checkpoint was loaded together with thresholds and span-gap parameter that showed best performance on the validation set. The model was applied on the test set for prediction with no further parameter tuning.

The token level predictions with confidence scores were then converted to binary predictions using category specific confidence thresholds. The character spans were reconstructed with the span merging procedure, specified by the span-gap parameter, and finally, spans along with their hallucination labels and confidence scores were saved in the required JSON file formatted for submission.

4 Experimental Setup

The proposed system was implemented in Python 3.12.3 using the PyTorch 2.6.0+cu124 deep learning framework and the Hugging Face Transformers library (version 5.12.1). The pretrained Qwen3-VL-8B-Instruct model was employed as the vision-language backbone. Image preprocessing was performed using the Pillow library, while NumPy was used for numerical computations. Scikit-learn was

used for auxiliary evaluation, including precision-recall analysis and classification reports. Additionally, official validation metrics (Cor and IoU) were calculated by applying the shared-task evaluation script (scorer.py) provided by the organizers. All experiments were conducted on an NVIDIA Tesla V100-PCIE-32GB with CUDA 12.2 support.

5 Results

5.1 Shared Task Results

Table 1 presents the official evaluation results of our submitted system on the hidden multilingual test set. The system is evaluated using three metrics: Cor+Lbl, Cor, and IoU, described further in Appendix A.2. Our approach has achieved 18th, 16th, 16th and 15th rank in English, French, Italian and Chinese language respectively.

Language	Cor+Lbl	Cor	IoU
English	0.2451	0.3108	0.2549
French	0.2429	0.3081	0.2672
Italian	0.2740	0.3784	0.3197
Chinese	0.3298	0.4172	0.3923

Table 1: Evaluation Result as per the organizers

5.2 Ablation Study on Classifier Head

Table 2 shows an ablation study that compares a simple linear token classification head and the proposed Conv1D span coherence head using the same training parameters. The results demonstrate that the Conv1D span coherence head achieves better performance than the baseline linear head in both metrics, IoU increases from 0.2967 to 0.3430 and Correlation increases from 0.3462 to 0.3910. This suggests that, on our experimental setup, modeling local contextual information using one-dimensional convolution tended to generate more coherent token level predictions, thus obtaining more accurate hallucination spans.

Classification Head	IoU	Cor
Linear	0.2967	0.3462
Conv1D	0.3430	0.3910

Table 2: Ablation study on different classification heads. Replacing the linear classification head with our proposed Conv1D span coherence head led to improvement in both IoU and Cor.

5.3 Ablation Study on Threshold Tuning

Figure 2 shows the evaluation results across the four languages before and after threshold tuning

Category	Avg. Annotation Probability	Fully Agreed (%)	Tuned Threshold	Δ (Tuned-0.50)
Invention	0.450	6.5	0.40	-0.10
Mischaracterization	0.428	5.0	0.40	-0.10
Miscounting	0.563	22.2	0.90	+0.40
OCR	0.551	16.6	0.65	+0.15
Other	0.368	0.8	0.15	-0.35

Table 3: Comparison of the average annotator agreement probability, the proportion of unanimously annotated spans, and the optimized category-specific decision thresholds.

and span merging.

Italian had the largest gain, where Cor increased by 0.036 (from 0.367 to 0.403) and IoU by 0.030 (from 0.319 to 0.349). English also showed relatively large gains of +0.022 (Cor) and +0.015 (IoU). The improvements for French were smaller, at +0.013 (Cor) and +0.017 (IoU). Finally, Chinese showed the lowest gains, increasing Cor and IoU by +0.008 and +0.010, respectively, as it already had the best performance among the languages at the baseline.

Threshold tuning and span merging increased the mean Spearman Correlation by +0.019 (+5.1%) from 0.375 to 0.394, and the mean IoU by +0.018 (+5.4%) from 0.333 to 0.351.

6 Conclusion & Discussion

We combined samples from all four languages into a single multilingual training set without language-specific sampling, weighting, or fine-tuning to test whether a single frozen Qwen3-VL-8B with a lightweight classification head would be able to generalize across languages and have a sufficiently simple deployment pipeline.

Our experiments demonstrate that freezing an LVLm and only training a small token classifier is a good approach for multilingual hallucination detection. The convolution classification head enables the model to learn local dependencies between neighboring tokens and, therefore, results in less noisy span predictions than an independent token classifier.

Moreover, the post-processing stage significantly impacted the final performance. Due to the contiguous nature of hallucinated spans, merging close spans decreases the artificial fragmentation of predicted hallucination spans, thus aligning more closely to the metric of the shared task.

Analysis of the annotation probabilities, as presented in Table 3, shows differences in annotator consensus across hallucination categories. Mis-

counting and OCR exhibit the highest average agreement (0.56 and 0.55), while the Other category has the lowest (0.37). We observe that the tuned decision thresholds learned during calibration follow the same ordering, with high-agreement categories requiring substantially higher thresholds (0.90 and 0.65) and ambiguous categories benefiting from lower thresholds (0.15). This suggests that category-specific calibration partly compensates for differences in annotation certainty encoded by the dataset’s soft labels.

Finally, the variation in the resulting scores between languages might be explained by language dependent annotation properties, tokenizer segmentation specificities, and different token level prediction and character level evaluation spans. Systematic investigation of these factors is an important direction for future research.

Limitations

The backbone of Qwen3-VL-8B-Instruct was kept frozen during the experiments. While this choice reduced the model’s demands on computational resources and memory, it limited the model’s ability to adapt its underlying multimodal representations to the specific characteristics of the task at hand. The thresholds were determined experimentally on the multilingual validation set; therefore, their generalizability on unseen data distributions and languages cannot be guaranteed. Lastly, the results showed that the model’s performance differed across languages. In particular, English and French scores were lower than Italian and Chinese scores, hinting at the possible benefits of tuning the model for each language separately.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhi-fang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng

Li, and 45 others. 2025. [Qwen3-vl technical report](#). *Preprint*, arXiv:2511.21631.

Yishuo Cai, Renjie Gu, Jiaxu Li, Xuancheng Huang, Junzhe Chen, Xiaotao Gu, and Minlie Huang. 2025. [MHALO: Evaluating MLLMs as fine-grained hallucination detectors](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9197–9222, Vienna, Austria. Association for Computational Linguistics.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 18135–18143.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. [Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models](#). *Preprint*, arXiv:2301.12597.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual instruction tuning](#). *Preprint*, arXiv:2304.08485.

Timothee Mickus, Claudio Savelli, Eduardo Calò, Emilio Raimond, Stella Frank, Hengyu Luo, Flavio Giobergia, Vincent Segonne, Chuyuan Li, Aman Sinha, Lorenzo Vaiani, Jörg Tiedemann, and Raúl Vázquez. 2026. [Can humans dream of electric sheep? human-written samples for fine-grained vision-and-language hallucination benchmarking](#). *Preprint*, arXiv:2608.01021.

Nusrat Jahan Prottasha, Upama Roy Chowdhury, Shetu Mohanto, Tasfia Nuzhat, Abdullah As Sami, Md Shamol Ali, Md Shohanur Islam Sobuj, Hafijur Raman, Md Kowsher, and Ozlem Ozmen Garibay. 2025. [Peft a2z: Parameter-efficient fine-tuning survey for large language and vision models](#). *Preprint*, arXiv:2504.14117.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2019. [Object hallucination in image captioning](#). *Preprint*, arXiv:1809.02156.

Raúl Vázquez, Aman Sinha, Chuyuan Li, Artem Shelmanov, Artem Vazhentsev, Claudio Savelli, Eduardo Calò, Emilio Raimond, Stella Frank, Hengyu Luo, Flavio Giobergia, Vincent Segonne, Lorenzo Vaiani, Jörg Tiedemann, and Timothee Mickus. 2026. [Overview of shroom-vision 2026: A shared task on hallucination detection in large vision-language models](#). *Preprint*, arXiv:2608.25662.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. 2024a. [Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation](#). *Preprint*, arXiv:2311.07397.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024b. [Mitigating hallucinations in large vision-language models with instruction contrastive decoding](#). *Preprint*, arXiv:2403.18715.

A Appendix

A.1 Classifier Head

Component	Type	Config.	Parameters
Dropout	nn.Dropout	p = 0.2	0
Span-coherence conv (weight)	nn.Conv1d	in = 4096 out = 4096 kernel = 5 padding = 2	83,886,080
Span-coherence conv (bias)	nn.Conv1d	out = 4096	4,096
Activation Function	nn.GELU	–	0
Output projection (weight)	nn.Linear	in = 4096 out = 5	20,480
Output projection (bias)	nn.Linear	out = 5	5
Total trainable			83,910,661

Table A.1: Configuration of the proposed token-level classification head and the number of trainable parameters for each component.

A.2 Metrics Used

- **Spearman Correlation (Cor)** measures the Spearman rank correlation between the predicted and reference per-character confidence vectors for each sample. The reference confidence values are the annotated soft labels (*prob*), while the predicted confidence values are obtained by projecting token-level predictions to character-level confidence scores.
- **Label-wise Spearman Correlation (Cor+Lbl)** extends the Spearman correlation by computing it separately for each hallucination category present in either the reference or the prediction for a sample. The final score is obtained by averaging these label-wise correlations, rewarding predictions that correctly identify both the confidence values and the hallucination categories.
- **Intersection over Union (IoU)** measures the overlap between the predicted and reference hallucination spans at the character level. A higher IoU indicates more accurate localization of hallucinated spans, independent of the predicted confidence scores.