

Team HalluciNauts at SHROOM-Visions: Anchor-Preserving Dual-Resolution Grounding for Fine-Grained Vision Hallucination Detection

Khang Nhat Hoang Vo

Mohamed bin Zayed University of Artificial Intelligence

Correspondence: Khang.Vo@mbzuai.ac.ae

Abstract

We describe our submission to the SHROOM-Visions 2026 English task, which requires detecting hallucinated character spans in image-conditioned responses, assigning confidence scores, and classifying each span as invention, mischaracterization, OCR error, miscounting, or other. Our system combines ModernBERT-large with frozen SigLIP2 patch features at 224×224 and 384×384 resolutions. A primary cross-modal detector produces token-level predictions, while a high-resolution branch adds bounded grounding and anchor-preserving residual corrections. The system uses only the official English training data and publicly available pretrained backbones, without retrieval, generated captions, external OCR, or additional labeled data. Post-hoc ablations show that the anchor-preserving correction provides the clearest incremental gain, while mismatched-image supervision has negligible development impact and histogram calibration is not beneficial. The submitted system achieves 0.2942 COR and 0.2140 COR+LBL on the public leaderboard. Our implementation is available here: <https://github.com/nhatkhangcs/Shroom-EMNLP-26.git>.

1 Introduction

Multimodal large language models can generate fluent responses containing objects, attributes, relations, text, or quantities unsupported by the input image. Early work evaluated this behavior mainly through hallucinated object mentions (Rohrbach et al., 2018; Li et al., 2023), while later studies considered finer-grained entity, attribute, relation, and reasoning errors (Gunjal et al., 2024; Cai et al., 2025). Because a response may be largely correct while containing only a short unsupported phrase, hallucination detection must localize the erroneous text rather than assign a single response-level label.

MHALO studies token-level multimodal hallucination detection (Cai et al., 2025), whereas

Mu-SHROOM formulates textual hallucination detection as character-level span prediction with soft annotator-derived probabilities (Vazquez et al., 2025). SHROOM-Visions 2026 extends this setting to image-conditioned generation: systems receive an image, prompt, and generated response and predict hallucinated character spans, confidence scores, and one of five error classes (Vázquez et al., 2026; Mickus et al., 2026).

Our output-only baseline captured textual hallucination patterns but could not verify claims against the image. Converting images into generated evidence yielded only a small class-aware gain and introduced parsing errors, motivating direct fusion with visual patch features.

Our final system combines ModernBERT-large (Warner et al., 2025) with frozen SigLIP2 features (Tschannen et al., 2025). A primary 224×224 branch produces the initial predictions, while a 384×384 branch supplies high-resolution token-image information for bounded residual refinement. Our main architectural contribution is an anchor-preserving correction stage that can modify hallucination and class logits without freely perturbing the primary detector. We additionally report component and loss-weight ablations to distinguish useful refinements from components that did not produce measurable development gains. The submitted system achieves 0.2942 COR and 0.2140 COR+LBL on the public leaderboard.

2 Related Work

Early visual hallucination benchmarks focused primarily on unsupported object mentions. CHAIR evaluates hallucinated objects in image captions (Rohrbach et al., 2018), while POPE tests object presence through binary questions (Li et al., 2023). Later work considers finer-grained entity, attribute, relation, and reasoning errors. M-HalDetect provides fine-grained annotations for

image-conditioned responses (Gunjal et al., 2024), MHALO studies token-level multimodal hallucination detection (Cai et al., 2025), and Mu-SHROOM formulates hallucination detection as character-level span prediction with soft annotator-derived probabilities (Vazquez et al., 2025).

Recent work has also investigated visual grounding signals for localized hallucination detection. VADE models patterns in a VLM’s internal visual-attention maps for token-level prediction (Prabhakaran et al., 2025), while Nguyen et al. (2026) analyze token-to-patch grounding across layers and combine grounding statistics with hidden representations. Our approach differs in that the detector does not read attention or hidden states from the VLM that generated the response. Instead, ModernBERT response representations explicitly attend to frozen SigLIP2 patch features, and a separate high-resolution pathway contributes bounded residual corrections to the supervised hallucination logits. This separates response encoding from visual evidence extraction and constrains later refinement around an already trained detector.

3 Methodology

Our system maps an image, a prompt, and a generated response to hallucinated character spans, confidence scores, and hallucination classes. As illustrated in Figure 1, ModernBERT-large encodes the prompt and response, while two frozen SigLIP2 encoders represent the image at different resolutions. The primary cross-modal branch predicts token-level hallucination and class logits, auxiliary span boundaries, and record-level hallucination presence. A high-resolution branch produces two bounded residual updates to the hallucination and class logits. The resulting token probabilities are fused with the record-level prediction, calibrated, and decoded into character spans.

3.1 Data and Preprocessing

We use the official English training partition of 3,799 responses associated with 1,607 unique images. An image-grouped 85/15 split yields 3,228 training and 571 validation records with no image overlap between partitions. Prompt and response text are retained unchanged, response offsets are preserved for character-level projection, and SigLIP2 patch features are cached before training. We use no retrieval, generated captions, external OCR, or additional labeled data.

3.2 Input Encoding

The prompt and generated response are tokenized as a sequence pair and encoded by ModernBERT-large (Warner et al., 2025). The response is retained as the second sequence so that its token offsets can be mapped independently to the original response string. The encoder produces contextual token representations $\mathbf{H}_{\text{text}} = [\mathbf{h}_1, \dots, \mathbf{h}_T] \in \mathbb{R}^{T \times H}$ with a maximum sequence length of 2,048 tokens.

The image is processed independently by two frozen SigLIP2 encoders (Tschannen et al., 2025). The 224×224 encoder produces $\mathbf{V}^{224} \in \mathbb{R}^{196 \times 768}$, while the 384×384 encoder produces $\mathbf{V}^{384} \in \mathbb{R}^{576 \times 768}$. Both sets of patch features are linearly projected to the ModernBERT hidden dimension.

3.3 Dual-Resolution Cross-Modal Fusion

For each resolution $r \in \{224, 384\}$, the ModernBERT token representations serve as queries, while the projected image patches serve as keys and values in an eight-head cross-attention module (Vaswani et al., 2017):

$$\mathbf{a}_i^r = \text{MHA}(\mathbf{h}_i, \tilde{\mathbf{V}}^r, \tilde{\mathbf{V}}^r). \quad (1)$$

We construct an explicit token–image interaction representation as

$$\mathbf{z}_i^r = [\mathbf{h}_i; \mathbf{a}_i^r; \mathbf{h}_i \odot \mathbf{a}_i^r; |\mathbf{h}_i - \mathbf{a}_i^r|]. \quad (2)$$

The concatenation retains the textual and attended visual representations, while the element-wise product and absolute difference expose their agreement and disagreement. The interaction is projected as $\hat{\mathbf{z}}_i^r = \text{GELU}(\mathbf{W}_f^r \mathbf{z}_i^r + \mathbf{b}_f^r)$, with gate $\mathbf{g}_i^r = \sigma(\mathbf{W}_g^r \mathbf{z}_i^r + \mathbf{b}_g^r)$. The fused representation is given by

$$\tilde{\mathbf{h}}_i^r = \text{LayerNorm}(\mathbf{h}_i + \mathbf{g}_i^r \odot \hat{\mathbf{z}}_i^r). \quad (3)$$

The two branches use independent visual projections, cross-attention modules, and fusion gates.

3.4 Primary Prediction Heads

The fused 224×224 representation $\tilde{\mathbf{h}}_i^{224}$ is passed to the primary prediction heads. For every response token, the model predicts a hallucination logit $\ell_i^{h, \text{base}}$ and five class logits $\ell_i^{c, \text{base}}$. It also predicts auxiliary start and end boundary logits and one record-level hallucination logit.

We use class-specialized heads instead of a shared classifier. Invention and mischaracterization use separate heads for object-existence and

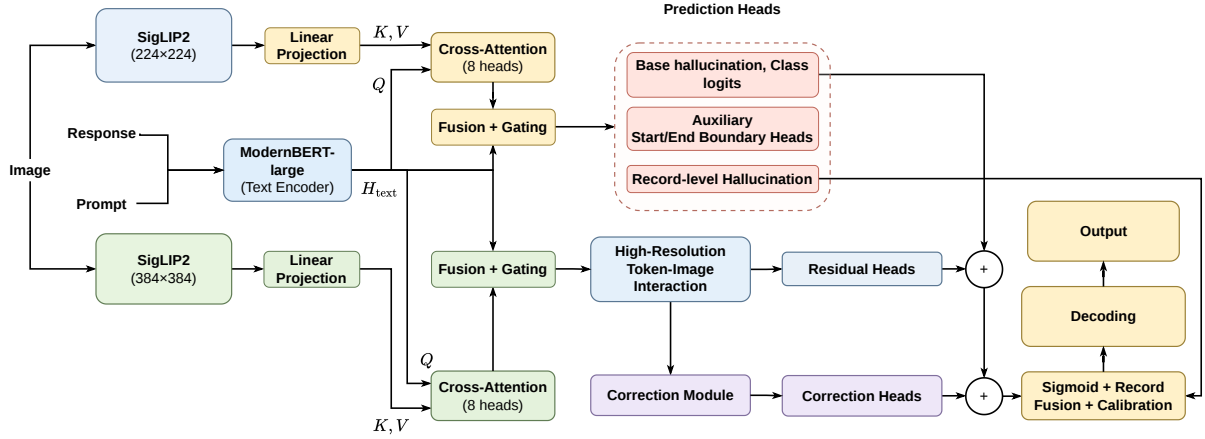


Figure 1: Overview of the proposed system. ModernBERT representations attend independently to frozen SigLIP2 patch features at two resolutions. The primary branch predicts hallucination and class logits, auxiliary boundaries, and record-level hallucination presence. The high-resolution interaction is processed by grounding and correction modules that add bounded residuals to the token-level hallucination and class logits. The final predictions are fused with the record-level score, calibrated, and decoded into character spans.

attribute–relation evidence. OCR and miscounting share a visually oriented head, while *other* uses a text-oriented head.

The boundary heads provide auxiliary supervision during training and are not modified by the high-resolution residual branches. The record-level prediction also bypasses the residual additions and is fused with the final token-level probabilities during calibration.

3.5 High-Resolution Residual Refinement

The fused 384×384 representation is combined with the original ModernBERT state to form the high-resolution interaction used by both residual branches for refinement:

$$\mathbf{q}_i = \left[\mathbf{h}_i; \tilde{\mathbf{h}}_i^{384}; \mathbf{h}_i \odot \tilde{\mathbf{h}}_i^{384}; \left| \mathbf{h}_i - \tilde{\mathbf{h}}_i^{384} \right| \right]. \quad (4)$$

Grounding residual. A grounding adapter maps $\mathbf{q}_i$ to hallucination and class residuals $\delta_i^{h,g}$ and $\delta_i^{c,g}$. Their magnitudes are bounded by applying $\tanh$ with scale $\rho_g = 0.35$. The output projections are initialized at zero, so this branch initially preserves the primary predictions.

To discourage reliance on response-only patterns, the grounding adapter is trained with mismatched-image negatives. For each response R_j , the corresponding image I_j^+ is contrasted against a randomly sampled image I_j^- :

$$\mathcal{L}_{\text{wrong}} = \frac{1}{B} \sum_{j=1}^B \max(0, m - s_\theta(I_j^+, R_j) + s_\theta(I_j^-, R_j)). \quad (5)$$

where $m = 0.25$ and s_θ is the learned image–response compatibility score.

Anchor-preserving correction. A second residual branch operates on the same high-resolution interaction. It projects $\mathbf{q}_i$ to a 256-dimensional correction representation using layer normalization, a linear layer, GELU activation, and dropout. The correction heads predict hallucination and class residuals $\delta_i^{h,r}$ and $\delta_i^{c,r}$, bounded with scale $\rho_r = 0.50$. Their output projections are also initialized at zero. The correction module contains 1,058,566 trainable parameters.

Only the token-level hallucination and class logits receive residual updates:

$$\begin{aligned} \ell_i^h &= \ell_i^{h,\text{base}} + \delta_i^{h,g} + \delta_i^{h,r}, \\ \ell_i^c &= \ell_i^{c,\text{base}} + \delta_i^{c,g} + \delta_i^{c,r}. \end{aligned} \quad (6)$$

The auxiliary boundary and record-level predictions remain those of the primary detector.

3.6 Soft Supervision and Optimization

For a response of C characters, we define a soft hallucination target $\mathbf{y}^h \in [0, 1]^C$, five class targets $\mathbf{Y}^c \in [0, 1]^{C \times 5}$, and start/end targets $\mathbf{Y}^b \in [0, 1]^{C \times 2}$. Each annotated span assigns its empirical probability to the covered characters and corresponding class channel. Overlapping annotations are summed and clipped to $[0, 1]$, preserving their combined evidence while maintaining valid soft targets. A record-level target indicates whether any hallucination is annotated.

Character-level targets are projected to response tokens using the maximum target value over the characters covered by each token. Tokens outside the response are excluded from token-level losses. The primary detector uses a learning rate of 2×10^{-5} and is selected at epoch 8. The grounding and correction modules use a learning rate of 10^{-4} and are selected at epoch 3.

Primary detector. The primary detector is trained with soft hallucination, class, Dice, boundary, and record-level objectives:

$$\mathcal{L}_{\text{primary}} = \mathcal{L}_{\text{hall}} + 0.75\mathcal{L}_{\text{class}} + 0.25\mathcal{L}_{\text{Dice}} + 0.15\mathcal{L}_{\text{boundary}} + 0.35\mathcal{L}_{\text{record}}. \quad (7)$$

As shown in Section 4.2, a simplified hallucination-and-class objective performs better in official-test ablation, indicating that the auxiliary-loss weighting is not a robust optimum.

Grounding adapter. After training the primary detector, its parameters are frozen. The grounding adapter is optimized using token-level hallucination and class supervision together with the wrong-image objective in Equation 5.

Correction module. The primary and grounding components are then frozen and treated as an anchor. Let $\bar{p}_i$ denote the anchor hallucination probability, p_i the corrected probability, and y_i the gold soft target. The agreement weight is defined inline as $w_i = \max(0.05, 1 - |\bar{p}_i - y_i|)$. Tokens on which the anchor agrees with the target are therefore preserved more strongly, while anchor errors retain greater freedom to change.

The hallucination consistency loss is

$$\mathcal{L}_{\text{anchor}}^h = \frac{\sum_i w_i (p_i - \bar{p}_i)^2}{\sum_i w_i}. \quad (8)$$

Records without annotated hallucinations receive three times the normal consistency weight. An analogous class-consistency loss is computed over gold hallucinated regions.

The correction module is trained with

$$\mathcal{L}_{\text{correction}} = \mathcal{L}_{\text{hall}} + 0.75\mathcal{L}_{\text{class}} + 0.25\mathcal{L}_{\text{Dice}} + 100\mathcal{L}_{\text{anchor}}^h + 20\mathcal{L}_{\text{anchor}}^c. \quad (9)$$

The anchor coefficients were selected empirically. Official-test ablations in Section 4.2 show that removing consistency degrades all three metrics, while increasing the coefficients beyond 100/20 provides no additional benefit.

System	Cor	Cor+Lbl
Output-only token tagger	0.2066	0.1155
Generated-evidence baseline	0.2026	0.1238
Dual-resolution residual detector	0.2942	0.2140

Table 1: Public leaderboard results for SHROOM-Visions 2026 English.

3.7 Calibration and Span Decoding

The submitted decoder passes token-level hallucination logits through a sigmoid and fuses them with the record-level probability using weight 0.35. The resulting scores use a 100-bin hallucination calibrator, while each class uses an independent 25-bin calibrator fitted on validation data. Section 4.2 shows that this calibration did not improve development performance.

A response token is activated when its calibrated hallucination probability exceeds 0.25. The global class threshold is 0.30 with a class margin of 0.15. The class-specific thresholds are 0.50 for invention and other, and 0.30 for mischaracterization, OCR, and miscounting. Token predictions are projected to characters through the tokenizer offset map. Contiguous characters assigned the same class form a span, and spans separated by at most three non-alphanumeric characters are merged. Final span confidence is the mean hallucination probability over its active characters.

4 Experimental Results

4.1 Official Results

COR measures correlation with empirical character-level hallucination probabilities, while COR+LBL additionally evaluates the assigned hallucination classes. Table 1 shows that the final multimodal detector substantially outperforms both early text-based baselines. The submitted system reaches 0.2942 COR and 0.2140 COR+LBL on the public leaderboard and obtains 0.3119 and 0.2570, respectively, on the grouped development split.

4.2 Component and Sensitivity Analysis

Table 2 reports controlled ablations on the official test set. The architectural sequence shows modest but consistent gains from the additional visual pathway: COR+LBL increases from 0.2115 for the 224×224 detector to 0.2122 with high-resolution grounding and 0.2140 after anchor-preserving correction. The corresponding COR scores increase from 0.2906 to 0.2920 and 0.2942. Thus, the high-

Variant	Cor+Lbl	Cor
<i>Architecture</i>		
224 × 224 only	.2115	.2906
+384 × 384 grounding	.2122	.2920
+ anchor correction	.2140	.2942
<i>Anchor consistency</i>		
No consistency	.2118	.2920
Reported 100/20	.2146	.2933
Strong 400/80	.2141	.2931
<i>Primary objective</i>		
Hallucination + class	.2234	.2839
Reported weights	.2180	.2800

Table 2: Official-test ablations of the principal architectural and loss choices.

resolution branch alone provides only a small gain, while the correction stage gives the clearest incremental architectural improvement.

Anchor-consistency ablations provide stronger evidence for the correction mechanism. Removing consistency reduces COR+LBL from 0.2146 to 0.2118 and COR from 0.2933 to 0.2920. Increasing the anchor coefficients from the reported 100/20 to 400/80 gives no further improvement, indicating that sufficiently strong anchoring is useful but that larger coefficients are unnecessary. The primary-loss ablation gives a different result. Training only with hallucination and class supervision reaches 0.2234 COR+LBL and 0.2839 COR, compared with 0.2180 and 0.2800 for the reported auxiliary-loss weights.

5 Conclusion

We presented a dual-resolution detector for fine-grained hallucination localization in SHROOM-Visions 2026. The submitted system combines ModernBERT response representations with frozen SigLIP2 patches and bounded high-resolution residual refinement, obtaining 0.2942 COR and 0.2140 COR+LBL on the English leaderboard. Additional ablations show that the components contribute unequally. High-resolution grounding alone provides only a modest improvement, while anchor-preserving correction gives the clearest architectural gain and strong anchor consistency improves class-aware correlation. Conversely, mismatched-image supervision yields no measurable development benefit and post-hoc experiments favor raw sigmoid scores over the submitted calibration.

Limitations

The annotations identify hallucinated spans but not the image regions that justify them, so cross-modal attention should not be interpreted as explicit or causal grounding. Frozen SigLIP2 features may also remain limited for OCR, fine-grained counting, small attributes, and spatial relations. The mismatched-image objective shows no measurable development benefit, and post-hoc experiments favor raw sigmoid scores over the submitted histogram calibration. We therefore do not attribute performance gains to these components. Finally, our experiments do not establish superiority over alternative grounding-based detectors such as VADE or [Nguyen et al.](#); controlled comparisons across common backbones and datasets remain in our future work.

References

- 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.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.
- 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.
- Tuan Dung Nguyen, Minh Khoi Ho, Qi Chen, Yutong Xie, Nguyen Cam-Tu, Minh Khoi Nguyen, Dang Huy Pham Nguyen, Anton van den Hengel, Johan W. Verjans, Phi Le Nguyen, and Vu Minh Hieu Phan. 2026. [Beyond the global scores: Fine-grained token grounding as a robust detector of lvlm hallucinations](#). *Preprint*, arXiv:2604.04863.
- Vishnu Prabhakaran, Purav Aggarwal, Vinay Kumar Verma, Gokul Swamy, and Anoop Saladi. 2025.

- VADE: Visual attention guided hallucination detection and elimination. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14949–14965, Vienna, Austria. Association for Computational Linguistics.
- Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.
- Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. 2025. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.
- Raul Vazquez, Timothee Mickus, Elaine Zosa, Teemu Vahtola, Jörg Tiedemann, Aman Sinha, Vincent Segonne, Fernando Sanchez-Vega, Alessandro Raganato, Jindřich Libovický, Jussi Karlgren, Shaoxiong Ji, Jindřich Helcl, Liane Guillou, Ona De Gibert, Jaione Bengoetxea, Joseph Attieh, and Marianna Apidianaki. 2025. SemEval-2025 task 3: MUSHROOM, the multilingual shared-task on hallucinations and related observable overgeneration mistakes. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2472–2497, Vienna, Austria. Association for Computational Linguistics.
- 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 Viani, Jörg Tiedemann, and Timothee Mickus. 2026. [Overview of shroom-visions 2026: A shared task on hallucination detection in large vision-language models](#). *Preprint*, arXiv:2608.25662.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

A System’s Final Rank

Our submitted system, under the team name HALLUCINAUTS, ranked 23rd among 28 leaderboard entries according to the primary COR+LBL metric. It achieved a COR+LBL score of 0.2140 and a COR score of 0.2942.

Team	Rank	Cor+Lbl	Cor
HALLUCINAUTS	23	0.2140	0.2942
Official baseline	27	0.1549	0.1549

Table 3: Final public leaderboard position of our submitted system. Rankings are ordered by COR+LBL.