

L1cache at SHROOM-Visions: MycoGate-VL for Scorer-Aware Multilingual Hallucination Detection

Aman Behera

Indian Institute of Technology Roorkee

aman_b@ch.iitr.ac.in

Abstract

SHROOM-Visions 2026 asks systems to locate and categorize hallucinated spans in multilingual vision-language responses. We describe MycoGate-VL, which ranks characters with a mainly text-based multilingual detector ensemble, labels them with a vision-trained category head, abstains on likely clean documents with a learned gate, and orders its output spans to match how the released scorer rebuilds character probabilities. On the official final leaderboard our per-language submissions reach mean COR .378 and COR+LBL .274, about .2 below the top systems; no ZH IoU is available because that submission’s predictions were not available for the IoU scoring added later. Across our leaderboard submissions, most of the COR+LBL rise from .05 to .25 arrived with changes to output construction (per-category spans, clean-document routing, and span ordering), although these steps were bundled with model updates. The ordering step alone changes no model predictions. On our held-out split the system beats the mark-nothing and kit baselines on COR in every language but not on COR+LBL for English and Chinese. Zero-shot uncertainty, judging, and retrieval did not improve performance in our small-scale pilots, whereas a vision-trained category head did. Our central finding is that output encoding and abstention policy can substantially affect performance alongside the underlying detector, although closing the gap to the top systems requires better detection.

1 Introduction

Vision-language models can produce fluent descriptions that are unsupported by the image (Ji et al., 2023; Li et al., 2023). SHROOM-Visions 2026 (Vázquez et al., 2026) evaluates this failure at character level: systems receive an image, a prompt, and a generated response, then predict hallucination probabilities and categories in English (EN),

French (FR), Italian (IT), and Chinese (ZH) (Figure 1). The data come from the SHEEP benchmark of human-written and model-generated samples (Mickus et al., 2026).

Compared with earlier SHROOM editions (Mickus et al., 2024; Vázquez et al., 2025), many documents are clean, and two clean rates matter. In our held-out split, 51–62% of documents have no span marked by a majority of annotators (majority-clean), and 25–36% have no span marked by any annotator (truly clean). The scorer rewards an empty prediction only on truly clean documents, so a useful system must rank suspicious text while knowing when to emit nothing. MycoGate-VL ranks characters with text-only multilingual encoders and uses the image only through a trained category head and, for EN and IT, a low-weight localization member; a learned clean-document gate and an output serializer built around the released scorer complete the system (Figure 2).

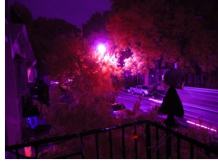
Our contributions are: (1) a span ordering that lets the scorer recover the detector’s own scores for COR and, at the same time, a separate ranking for each category for COR+LBL (§5); (2) a learned gate that abstains on likely clean documents, turning the scorer’s reward for empty output on clean documents into an explicit decision; and (3) small-scale negative results for zero-shot uncertainty, multimodal judging, and retrieval, alongside one supervised vision lever (a trained category head) that helped.

Code, the scorer replica, and training scripts are public.¹

2 Related Work

The SHROOM series established character-level detection with multi-annotator soft labels (Mickus et al., 2024; Vázquez et al., 2025). Many existing

¹<https://github.com/beingamanforever/MycoGate-VL>



input image

Prompt: How many lights are hanging to black fence?

Response: Based on the image, there are ****8**** lights hanging on the black fence. They appear to be string lights.

System output
probability for every character; one of five labels per span

gold for **8**: **miscounting** $\frac{2}{3}$ • **invention** $\frac{1}{3}$ • $\Pr(\text{hallucinated}) = 1$

Figure 1: A training example, reproduced verbatim with its gold annotation (document `train-en-641`; the asterisks are the generator’s own markdown emphasis). All three annotators marked the character “8” (the image shows more lights), so its gold hallucination probability is 1; two labelled it miscounting and one invention, giving the per-category fractions shown.

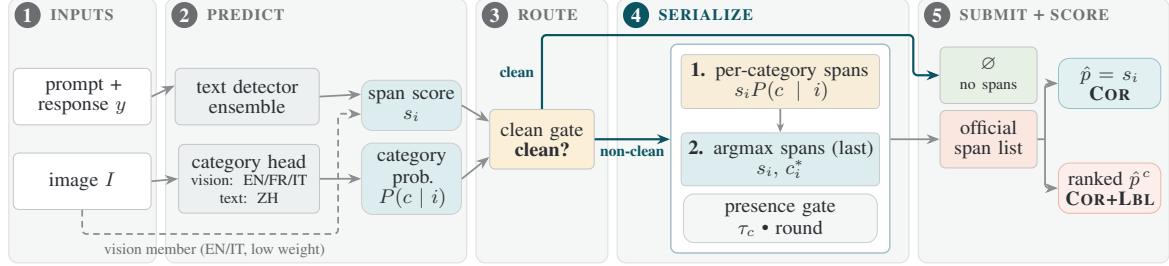


Figure 2: MycoGate-VL inference pipeline. Text detectors produce span scores and a text or vision head produces category probabilities; a dashed low-weight vision member contributes to localization for EN/IT only. The clean gate abstains on likely clean responses; otherwise the serializer writes per-category spans first and argmax spans last. Because the scorer lets the last span covering a character win, it recovers the unlabelled score vector for COR while still seeing per-category rankings for COR+LBL (worked example in Table 2).

LVLBM benchmarks focus on whether a response contains hallucinations (Guan et al., 2024; Wang et al., 2023; Yin et al., 2024); this task additionally requires exact localization and multilingual category labels. We test zero-shot uncertainty estimation (Fadeeva et al., 2023; Kuhn et al., 2023; Farquhar et al., 2024), retrieval, and multimodal judging, but our core remains token-level soft regression with fine-tuned multilingual encoders (Conneau et al., 2020; Marone et al., 2025).

3 Task and Evaluation

The task evaluates systems with character-level span overlap (IoU) and confidence-based correlation metrics (Vázquez et al., 2026). During evaluation, the participant kit and platform exposed the correlation scores as COR and its per-category extension COR+LBL. The organizers later added IoU rankings and reopened the platform for supplementary predictions; we treat the final leaderboard on the official task website as definitive. We report all three leaderboard rows and use COR and COR+LBL for the scorer-aware analysis below.²

²The official leaderboard was accessed August 9, 2026 and re-checked on September 11, 2026; its values match the scores we recorded at submission time. The leaderboard ta-

Annotators mark hallucinated character spans and assign one of five scorer labels: invention, mischaracterization, OCR, miscounting, or other.³ Their marks define the soft gold probability

$$\Pr(c_i) = \frac{1}{N_a} \sum_{n=1}^{N_a} \mathbf{1}[c_i \in C^{(n)}],$$

where N_a is the number of annotators and $C^{(n)}$ the character set annotator n marked; values are usually in $\{0, \frac{1}{3}, \frac{2}{3}, 1\}$. From submitted spans, the scorer rebuilds one probability $\hat{p}_i$ per character by writing the spans in list order, so where spans overlap, the last span covering a character wins (we call this *overwrite reconstruction*). It reports, per document,

$$\text{COR} = \text{Spearman}(\Pr, \hat{p}), \quad (1)$$

$$\text{COR+LBL} = \frac{1}{|\mathcal{K}_d|} \sum_{k \in \mathcal{K}_d} \text{Spearman}(\Pr^k, \hat{p}^k), \quad (2)$$

bles of the overview paper (Vázquez et al., 2026) give our team higher COR+LBL and COR on EN and FR (for example EN COR+LBL .267 against .2273 here) and agree within .01 elsewhere; we could not determine the cause and report the website values throughout. IoU values computed on held-out data are local diagnostics; the IoU values and ranks in Table 3 are official test-set results.

³The task page describes the labels as “one of four categories” in its introductory text but lists five categories. The gold data uses the label “OCR” for the third category; we therefore follow the five labels accepted by the scorer.

Approach	Outcome
White-box signals	Near-chance AUC on a 10-document pilot; model confidence may track fluency rather than what annotators marked.
Zero-shot judge	Uses the image, but remains well below the trained span detector (40-document test).
Wikipedia retrieval	No reliable gain; consistent with most errors being perceptual.
Supervised vision	Improves category labels, not primary span localization.

Table 1: Alternatives we evaluated; details and scales in Appendix Tables 8–10.

where $\mathcal{K}_d$ is the union of categories present in the reference or the prediction (the *union rule*), and $\Pr^k, \hat{p}^k$ are the per-category vectors. If either vector is constant, for example an empty prediction or a document with no gold span, the score is 1 if both are constant and agree on whether any hallucination is present, and 0 otherwise (the *constant-vector rule*; Vázquez et al., 2026). This creates three practical constraints: empty output is rewarded on truly clean documents, a category present in the prediction but absent from the gold adds a zero term, and only within-document ranks matter. In our held-out split, 51–62% of documents are majority-clean and 25–36% are truly clean (Appendix Table 11); only the latter earn the constant-vector reward, which makes over-prediction an important source of error. Our scorer replica is regression-tested on adversarial overlap, constant-vector, and category-union cases, and was sufficiently consistent for the structural effects reported here, but not for fine serialization deltas ($\pm .005$), for which the leaderboard was the arbiter.

4 What Helped and What Did Not

Three zero-shot alternatives did not yield improvements in our small-scale tests (Table 1), while supervised vision produced a narrower, useful gain.

White-box uncertainty signals (Qwen2.5-VL’s own token probabilities; Bai et al., 2025), multi-modal judging, and retrieval all scored below task-specific supervision in these tests. One plausible explanation is annotation alignment: gold labels record what annotators selected rather than every visually defensible error. Supervised vision is different. It did not improve localization consistently, but a vision-trained category head improved label prediction in EN, FR, and IT. We therefore use

the image selectively rather than throughout the system.

5 System

Detection and ensembling Four multilingual token regressors based on XLM-R and mmBERT (Conneau et al., 2020; Marone et al., 2025) predict a score s_i for every character; a second head predicts categories. The score s_i is trained to predict the gold probability of each character, and the submitted spans are derived from these scores alone by the serialization below. Where the loss up-weights hallucinated characters to counter their scarcity, we cap the weight at 4, because the uncapped weight hurt (Appendix Table 6a). We average complementary models trained from scratch and initialized from Mu-SHROOM. Following leaderboard feedback (Appendix Table 7), the final composite uses the ensemble for EN, a reweighted ensemble that adds the original detector as a fifth member for IT, and the original detector alone for FR and ZH. The image is used by the category head for EN, FR, and IT and as a low-weight localization member for EN and IT. Training details are in Appendix A; ablations in Appendix Table 6.

Scorer-aware serialization A submission is a list of spans, each with one label and one probability. The simplest encoding writes one span per run of characters with its argmax label and score s_i . This preserves COR, but it shows the scorer each category only on the characters where it is the argmax, so a category’s COR+LBL term can fall to zero or below. We instead write two layers: first, for every category c , spans with score $s_i P(c | i)$ (rounded to two decimals; values below .05 are dropped); second, the argmax spans with score s_i . Because the argmax spans are written last, overwrite reconstruction returns the detector’s own score vector (after rounding) for COR, while each per-category vector keeps that category’s ranking for COR+LBL. This is output encoding under the released scorer, not a change to model predictions.

Table 2 gives a worked example on the response “two red cats”: the detector scores “two” at .9 (miscounting .7, invention .3) and “red” at .6 (invention .6, miscounting .4). Suppose the gold marks “two” as miscounting ($\frac{2}{3}$) and invention ($\frac{1}{3}$), as in Figure 1. Both encodings score COR .82 under our replica, but COR+LBL is .54 for the two-layer list and .33 for the argmax-only list. The argmax-only list gives “two” an invention score of 0, so the invention term

(a) emitted span list, in writing order

#	span	label	prob.
<i>layer 1: per-category spans, $s_i P(c i)$</i>			
1	two	invention	.27
2	red	invention	.36
3	two	miscounting	.63
4	red	miscounting	.24
<i>layer 2: argmax spans, written last, s_i</i>			
5	two	miscounting	.90
6	red	invention	.60

(b) vectors the scorer rebuilds (last covering span wins)

Vector	two	red	rest
all labels, used for COR	.90	.60	0
miscounting, used for COR+LBL	.90	.24	0
invention, used for COR+LBL	.27	.60	0

Table 2: Worked example of the two-layer serialization on “two red cats” (spans given by word; all other characters are 0). The all-label vector equals the detector scores s_i , so COR is unchanged; each category vector keeps its own ranking, which an argmax-only list (spans 5–6) would lose.

is negative (−.33).

Routing and post-processing A learned per-language *clean gate* (a gradient-boosted classifier over document-level features of the text and the scores s_i) sends likely clean documents to an empty output; removing it causes the largest single-component drop we measured (Appendix Table 6a, ZH). On the remaining documents, a *presence gate* removes any category whose strongest evidence $\max_i s_i P(c | i)$ falls below a per-category threshold τ_c : its per-category spans are dropped, and its argmax spans are relabelled with the best remaining category, so the COR vector is unchanged (we check this on every document when files are built). This avoids the zero term that COR+LBL charges for a category present in the prediction but absent from the gold. Finally, scores are rounded: the per-category spans to two decimals, and the argmax spans to one decimal for EN and IT and two for FR and ZH (Appendix Table 5). Gold probabilities take at most $N_a + 1$ distinct values, so gold vectors contain many ties; Spearman correlation is highest when the prediction shares that tie structure, and rounding merges near-equal scores into ties (Appendix Table 6b; three decimals matched no rounding). All thresholds are selected per language against the scorer replica.

6 Results

We report two kinds of results. Leaderboard results (Table 3) come from files whose per-language variants and thresholds were chosen with leaderboard feedback, because the task allowed repeated and supplementary submissions. These results carry selection bias, and Table 3a lists cumulative best states, not controlled ablations (per-lever probes are in Appendix Table 7). Held-out results (Table 4) re-score the final per-language configuration on the 15% held-out split of the roughly 3,800 labelled documents per language (Appendix A), with the clean gate refit out of fold and equal ensemble weights (Appendix E). The split is disjoint from the test set, but the variant used for each language was still chosen with leaderboard feedback, so this is our controlled comparison against the baselines rather than a leaderboard-free selection. No peer rankings were visible to us during the original submission period. For context, the best systems reach mean COR .58, COR+LBL .46, and IoU .51 (Vázquez et al., 2026), about .2 above ours on each metric.

Output construction dominates the COR+LBL trajectory (Table 3a, rows 1–3). Per-category spans with the clean gate (row 2, which also retrained the detector) and the two-layer serialization (row 3, which changes no model predictions) accompany the rise from .05 to .25. In held-out tests, the two-layer serialization alone, compared with argmax-only spans from the same detector, increases COR+LBL from .11–.15 to .22–.27 at unchanged COR. On the leaderboard, ensembling mainly improved COR, while category and presence decisions improved COR+LBL (Appendix Table 7). In the held-out comparison (Table 4), our system leads on COR in all four languages, with paired bootstrap intervals that exclude zero (Appendix Table 11). On COR+LBL it leads B0 and the kit baseline for FR and IT but remains below both on EN and ZH, and none of the four COR+LBL differences from B0 is distinguishable from zero. The reason is structural: the union rule charges a zero term for every falsely present category, and B0 earns the constant-vector reward on every truly clean document. English member reweighting improved validation but regressed on the leaderboard, so the released system uses only leaderboard-confirmed per-language choices.

(a) cumulative leaderboard states, language mean

Milestone	Kind	COR+LBL	COR
argmax-only serialization	–	.048	.335
+ per-category spans & clean gate	O, M	.226	.357
+ two-layer serialization	O	.248	.356
+ rounding & per-language tuning	O	.255	.362
+ vision category head	M	.264	.362
+ ensemble & presence gate	M, O	.271	.376
best submission per language	M, S	.274	.378

(b) best submission per language: score (official rank)

	EN	FR	IT	ZH
COR+LBL	.2273 (20)	.2730 (15)	.2883 (15)	.3093 (16)
COR	.3344 (19)	.3637 (14)	.3977 (14)	.4152 (15)
IoU	.2557 (20)	.2931 (15)	.2843 (16)	N/A

Table 3: Official final leaderboard results for team L1cache. (a) Each row is the mean of the best leaderboard state per language after the named change landed; the kind column marks whether the step changed output construction (O), the model (M), or file selection (S). Adjacent rows can bundle auxiliary changes and per-language reversions, so row differences are not single-change effects; the last row adds the low-weight vision localization member for EN and IT and takes ZH from an earlier submission. (b) Per-language values behind the last row of (a): each cell gives our leaderboard score followed by the official rank for that metric and language. N/A denotes that the ZH submission’s predictions were not available for the IoU scoring the organizers added later, so no IoU score or rank exists; our IoU comparison therefore covers EN, FR, and IT only.

7 Discussion

MycoGate-VL shows that scorer semantics can shape system design: the detector ranks characters, the clean gate controls abstention, and the serialization preserves information that an argmax-only span list discards. Two caveats bound this finding. First, it holds in our score range; the top systems sit about .2 higher on every metric (Vázquez et al., 2026), most likely through stronger detection, and no re-encoding of our detector’s output closes that gap. Second, the serialization is specific to the span format. Under a protocol that consumed per-character, per-category probabilities directly, the two-layer serialization would be unnecessary: it exists only to carry the detector’s scores and per-category scores through the scorer’s span-list format. The presence gate is tied to the union rule of COR+LBL. The clean gate would remain useful under any metric that rewards abstention on clean documents. The detector itself is unchanged by all three. The zero-shot alternatives we tested remained poorly aligned with subjective span an-

Held-out	EN	FR	IT	ZH
COR				
mark-nothing (B0)	.253	.256	.263	.364
kit neural baseline	.289	.272	.279	.385
MycoGate-VL	.367	.369	.407	.433
COR+LBL				
mark-nothing (B0)	.253	.256	.263	.364
kit neural baseline	.255	.259	.261	.369
MycoGate-VL	.245	.269	.286	.335

Table 4: Held-out comparison (15% split, replica scorer with union COR+LBL). B0 predicts nothing; the kit baseline is the organizers’ LLaVA-1.5 token classifier (Appendix E). MycoGate-VL rows use an out-of-fold clean gate and equal-weight ensembling in place of the shipped per-language weights, so they are conservative estimates. Our system clears both baselines on COR in every language and on COR+LBL for FR and IT; on EN and ZH COR+LBL both baselines stay ahead of it; B0 collects the constant-vector reward on every truly clean document, which is worth most on ZH, where the clean rate is highest. Paired bootstrap intervals for the difference to B0 (Appendix Table 11) exclude zero for COR in every language and include zero for COR+LBL in every language.

notations in our pilots, whereas supervised vision helped the narrower category decision. Remaining errors are often boundary over-extensions (Appendix D); structured span heads are a natural next step.

8 Conclusion

Can careful output construction substitute for a stronger model on this task? Only partially, and only within our score range: under the released scorer, serialization and abstention decisions moved COR+LBL more than any modelling change we tried, and image information paid off only where it was trained against the gold labels (category prediction). The negative results for zero-shot uncertainty, judging, and retrieval come from small pilots and should be read as such. Reliable localization still required a supervised multilingual detector; treating ranking, category presence, and clean-document routing as separate decisions gave the best configuration we found.

Limitations

Held-out estimates come from a split of the training data; the test set may differ in clean rate, generator mixture, or annotator count, and the per-language operating points are tuned to the former. Because repeated leaderboard submissions

guided per-language file and threshold selection, the “best submission” figures carry an optimistic selection bias, and sequential submissions are not controlled experiments. Our held-out re-scoring (out-of-fold gate, union COR+LBL) shows our system behind both the mark-nothing and kit baselines on COR+LBL for EN and ZH, although the paired bootstrap intervals for the COR+LBL difference to the mark-nothing baseline include zero in all four languages, only marginally for ZH (Appendix Table 11); earlier drafts overstated this comparison. The test set also has fewer clean documents than our split: the overview paper reports that 15.5–25.5% of test documents carry no annotated span (Vázquez et al., 2026), against 25–36% truly clean documents in the split on which the gate thresholds were tuned. Our final composite files leave 28–55% of test documents empty (EN .34, FR .39, IT .28, ZH .55), so at least 12–29% of test documents were left empty although they carry an annotated span; we cannot measure the score cost because the test annotations are private. The localization scores are text-only: it detects the linguistic signature of hallucination rather than verifying against the image, which limits performance on hallucinations that require genuinely visual verification. We report no category-wise scores per language, so we cannot say for which categories (for example miscounting or OCR) the vision head helps most. The negative results are small-scale: the white-box study covers 10 Chinese training documents on one 7B VLM, and the judging results cover one 40-document English test (Gemini Flash) plus one 47-sentence Chinese probe (Qwen2.5-VL-7B); stronger or verification-tuned models could shift those conclusions. Validation predicted leaderboard direction reliably in Italian and French but not in English, so the English operating points are the least trustworthy part of the configuration. Category supervision is scarce for OCR and miscounting, and the synthetic augmentation was too small to test the boundary hypothesis it was built for. Finally, our replica and the kit scorer were not compared file-by-file on identical inputs, and they differ on one edge case, an empty prediction against a document whose every character is marked: the kit scorer compares only the number of distinct values on each side and scores it 1, while our replica follows the presence-versus-absence wording of the overview paper (Vázquez et al., 2026) and scores it 0. All held-out tables use the replica; the kit rule would change 3 of the 570 EN held-out documents

and none in FR, IT, or ZH, moving an EN cell by about .005 (B0 COR on EN would read .258 instead of .253).

Ethical Considerations

Hallucination detection is dual-use: it supports reliability and moderation, but a released detector could in principle guide evasion. We release the detector only, not a paired generator. Performance on languages beyond the four evaluated is untested and not claimed. Categorical predictions inherit any biases in the gold distribution, which we describe and do not correct. All training data is the official 2026 release; Figure 1 reproduces one training image and response verbatim from that release.

Acknowledgments

We thank the SHROOM-Visions organizers for the task and for clarifying the 2026 metrics and submission format.

References

- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuezhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. [Qwen2.5-VL technical report](#). *arXiv preprint arXiv:2502.13923*.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.
- Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. 2023. [LM-polygraph: Uncertainty estimation for language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 446–461. Association for Computational Linguistics.
- Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. [Detecting hallucinations in large language models using semantic entropy](#). *Nature*, 630:625–630.
- Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen,

- Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. [HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. [Survey of hallucination in natural language generation](#). *ACM Computing Surveys*, 55(12):1–38.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. [Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation](#). In *The Eleventh International Conference on Learning Representations (ICLR)*.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne 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. Association for Computational Linguistics.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. [Improved baselines with visual instruction tuning](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual instruction tuning](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations (ICLR)*.
- Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. [mmBERT: A modern multilingual encoder with annealed language learning](#). *arXiv preprint arXiv:2509.06888*.
- 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.
- Timothee Mickus, Elaine Zosa, Raúl Vázquez, Teemu Vahtola, Jörg Tiedemann, Vincent Segonne, Alessandro Raganato, and Marianna Apidianaki. 2024. [SemEval-2024 task 6: SHROOM, a shared-task on hallucinations and related observable overgeneration mistakes](#). In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1979–1993, Mexico City, Mexico. Association for Computational Linguistics.
- Raúl Vázquez, Timothee Mickus, Elaine Zosa, Teemu Vahtola, Jörg Tiedemann, Aman Sinha, Vincent Segonne, Fernando Sánchez-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 Vaiani, 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.
- Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. 2023. [AMBER: An LLM-free multidimensional benchmark for MLLMs hallucination evaluation](#). *arXiv preprint arXiv:2311.07397*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2024. [Woodpecker: Hallucination correction for multimodal large language models](#). *Science China Information Sciences*, 67(12):220105.

A Training Details

Table 5 lists the configuration of the text detectors, the serializer, and the gate, all trained with AdamW (Loshchilov and Hutter, 2019). The multilingual multitask trainer uses all rows as printed; scope notes mark rows that differ for other components. Data: 3,799/3,767/3,746/3,790 labelled documents (EN/FR/IT/ZH), split 85/15 into train/held-out per language (regime-stratified, seed 42); the held-out split also served as the validation set for threshold and variant selection.

encoders	xlm-roberta-base, mmBERT-base
max sequence length	512 (ZH: sliding windows)
batch size	8
optimizer	AdamW; lr 2×10^{-5} (LoRA: 10^{-4})
schedule	cosine, 10% warmup (multitask/ens.)
epochs	6 (multilingual), 4 (dedicated)
positive weight w^+	4 capped (multitask trainer); none (ensemble)
category loss weight λ	0.5
gradient clipping	1.0 (multitask/ensemble trainers)
split-layer floor	0.05
output rounding	2 dec.; argmax spans 1 dec. (EN, IT)
ensemble members	{XLM-R, mmBERT} $\times$ {scratch, warm}
category posterior	vision LoRA (Qwen2.5-VL); text: ZH
presence thresholds τ_c	per cat. & lang., grid .05–.5
clean gate	learned, truly-clean target
gate threshold g_ℓ	5-fold CV over grid .30–.99
hardware	NVIDIA L4 24 GB (detectors, kit baseline); L40/A40 48 GB (vision head, ensemble sessions)
seeds	42 (main); {42, 123, 999} (ensemble); 1 (gate CV); 0 (LoRA, tuning)

Table 5: Training and tuning configuration.

B Sanitization and Threshold Remap

Earlier single-span builds binarized the soft vector at a tuned per-language τ_ℓ and cleaned the resulting spans in codepoint space: leading and trailing punctuation and whitespace stripped by Unicode category (including full-width CJK punctuation), boundaries snapped to word edges (disabled for Chinese), zero-length and overlapping spans merged. To express a tuned $\tau_\ell \neq 0.5$ in a format whose hard set is the summed-probability majority, the strictly increasing remap $m_\tau(x) = x/2\tau$ for $x \leq \tau$, else $\frac{1}{2} + (x - \tau)/2(1 - \tau)$, places τ_ℓ at the scorer’s 0.5 point and, since only within-document ranks enter Eq. 1, leaves COR untouched. The two-layer serialization needs no binarization, so the final system keeps only the codepoint conventions.

C Ablation and Probe Tables

Table 6 reports the gate/loss and rounding ablations; Table 7 maps each lever to its leaderboard probe; Tables 8–10 detail negative results; Table 11 gives the document composition behind B0 and bootstrap intervals for the gap to B0.

D Boundary Error Example

A representative English validation case, gold majority versus our span; the detector anchors on the OCR misreading and absorbs the rest of the plausible-but-unmarked name (highlight colors mark the two span extents, labelled at left):

gold The name of the mall appears to be
 "SSTEIN Sentral Kuala Lumpur."
 ours The name of the mall appears to be
 "SSTEIN Sentral Kuala Lumpur."

(a) gate & loss weighting (ZH, 228-doc held-out)

Configuration	IoU [†]	COR
B0 (mark-nothing)	.649	.386
detector, no gate	–	.219
detector + gate (ours)	.652	.423
$w^+=36$ uncapped, no gate	.549	–

(b) Δ COR+LBL of two-decimal rounding vs. none

Comparison	EN	FR	IT	ZH
held-out sweep	+.004	+.011	+.008	+.008
leaderboard	–	+.008	+.009	+.008

Table 6: Ablations. (a) On a 228-document ZH held-out subsplit, the gate turns the same detector from losing to B0 on COR (.22) to winning (.42); the uncapped positive weight collapses the hard spans ($\text{IoU} .55 < \text{B0}$). Dashes mark cells whose original run logs were not preserved. (b) Two-decimal rounding helped on the held-out sweep in all languages and was leaderboard-confirmed for FR/IT/ZH (EN was not probed on the leaderboard); three-decimal rounding matched no rounding. [†]These IoU values are local held-out diagnostics, not official test-set leaderboard values (§3).

E Kit-Baseline Protocol and Reproducibility

Baseline reproduction The kit baseline teacher-forces LLaVA-1.5-7B (Liu et al., 2023, 2024) on each prompt–image–response and classifies tokens from probability, loss, hidden-state, and attention features (101 features in our replica of the kit’s featurizer). Current SDPA attention returns no attention maps in transformers (Wolf et al., 2020), so we use eager attention. Tokens are positive when their aligned character ranges overlap majority gold. Per language, we train on 1,500 documents (seed 42) and select logistic regression or a 300-tree random forest by F1 on a 10% token holdout; the forest wins in all four languages. Predictions use $p \geq 0.5$. 2 of 8,266 documents (both ZH) were skipped.

Reproducibility The replica mirrors the released scorer’s overwrite reconstruction and its 8-decimal constant-vector rule, and is regression-tested on adversarial overlap, constant-vector, and category-union cases (part of the repository’s 62-test suite). Seeds are 42 for data and classification and 1 for gate cross-validation. Presence gating preserves COR span by span (asserted per document at build time). Gate refits are deterministic for a fixed library version, but leaderboard-confirmed files ship verbatim rather than being rebuilt. Held-out system scores in Tables 4 and 11 use an out-of-fold

Leaderboard probe	EN	FR	IT	ZH
equal-weight ensembling	+.016	.000	+.033	+.002 ^a
member reweighting	−.011 ^b	−	+.009 ^c	−
presence gating	rej. ^b	+.011	+.029 ^c	− ^d
vision-blend member	+.001	rej. ^b	+.002	rej. ^b

Table 7: Per-lever leaderboard probes (differences between identified submission pairs; dashes: not probed). Rows report ΔCOR , except presence gating, which leaves COR unchanged by construction (§5) and is reported as $\Delta\text{COR}+\text{LBL}$. ^aMeasured but the earlier file was kept. ^bRejected: the change regressed on the leaderboard despite validation gains. ^cThe IT pair bundles reweighting with COR-neutral suppression (neutrality asserted per document at build time). ^dZH presence gating was tuned on validation (+.018) but the earlier labelled file was kept.

Signal (40 hallucinated EN docs)	COR	inv.
Trained text head	.237	−
Judge (Gemini Flash), verbatim spans	.060	.020
↪ <i>same, image removed</i> *	.013	−
Judge (Gemini Flash), dense clauses	.058	−

Table 8: Zero-shot multimodal judging versus the trained head (shaded) on hallucinated documents, where ranking matters. “inv.” is the per-category invention correlation, undefined for the soft head (−). *Derived from the verbatim variant’s measured image ablation (−.047); the dense variant’s image ablation was negative (removing the image *helped*).

clean gate (5 folds, seed 1) and equal-weight ensembling, regenerated by `experiments/score_heldout_oof.py` from our stored held-out predictions (outputs recorded in `docs/audit/heldout_rescore.json`); the paired bootstrap of Table 11 is produced by `experiments/heldout_bootstrap.py`. These cells require the LightGBM gate backend (reproduced with LightGBM 4.7.0); the repository’s numpy fallback trains a different gate and does not reproduce them.

F Judge Prompts

Both prompts below were run with Qwen2.5-VL-7B-Instruct on a 47-sentence probe over the 10-document Chinese pilot; the contradiction framing reached a sentence-level AUC of 0.646 but inverts at the character level, and the support framing over-flags non-visual content (AUC 0.484). The verbatim-span prompt is the Gemini Flash configuration of Table 8.

Support prompt. “*Is every detail in this statement directly supported by what is visible in*

White-box feature	AUC
NLL	.482
Entropy	.479
Image-ablation Δ	.492
Claim-conditioned prob.	.518
Max. probability	.545
<i>chance</i>	.500

Table 9: Zero-shot white-box features from teacher-forcing Qwen2.5-VL-7B: per-character AUC for predicting majority-hallucinated characters, on a pilot of 10 Chinese *training* documents (§4). At this sample size no feature separates from chance; a planned full-scale rerun was not executed.

RAG – text-only	$\Delta\text{IoU}^\dagger$	ΔCOR
EN	+.009 [−.006, +.024]	+.001 [−.019, +.023]
FR	−.003 [−.017, +.011]	−.005 [−.025, +.012]
IT	−.025 [−.049, −.0004]	−.016 [−.036, +.007]
ZH	−.008 [−.026, +.009]	+.002 [−.015, +.017]

Table 10: Wikipedia retrieval versus text-only (§4). Brackets are 95% paired per-document bootstrap CIs. No gain in seven of eight cells; the Italian IoU interval excludes zero, but only marginally (upper bound -4×10^{-4}) and without correction for the eight comparisons. [†]These IoU values are local held-out diagnostics; official test-set IoU is ranked on the final leaderboard.

the image? Answer Yes or No.”

Contradiction prompt. “*Does the image contradict this statement? If it is an opinion, a guess, or not about something visible, answer No. Answer Yes or No.*”

Verbatim-span prompt. “*Describe what is visible in the image. Then quote every unsupported span from the response verbatim and assign each span a category.*”

Held-out split	EN	FR	IT	ZH
majority-clean (%)	50.7	52.7	52.3	61.6
truly clean (%)	25.3	25.6	26.3	36.4
COR, B0	.253	.256	.263	.364
COR, ours	.367	.369	.407	.433
COR+LBL, B0	.253	.256	.263	.364
COR+LBL, ours	.245	.269	.286	.335
Δ COR	+.114	+.113	+.144	+.068
95% CI	[.07, .15]	[.07, .16]	[.10, .19]	[.03, .11]
Δ COR+LBL	-.007	+.012	+.023	-.030
95% CI	[-.04, .03]	[-.03, .05]	[-.01, .06]	[-.06, .003]

Table 11: Document composition, held-out scores, and the gap to B0 (same split and protocol as Table 4). Under the replica’s constant-vector rule, an empty prediction scores 1 exactly when the gold vector is all-zero, so B0’s COR and COR+LBL equal the truly-clean rate. Ours (shaded) trades some of that reward for ranking the rest; on COR+LBL the trade does not pay on EN/ZH. The Δ rows give ours minus B0 with paired per-document bootstrap intervals (10,000 resamples, seed 0; 562–570 documents per language): every COR interval excludes zero; no COR+LBL interval does, although the ZH upper bound is only .003. Only aggregate scores are available for the kit baseline.