

USP at SHROOM-Visions: Why Informative Proxy Visual Signals Do Not Localize Hallucinated Spans

Rafael Macário Fernandes

Department of Linguistics

University of São Paulo

rafael.macario@usp.br

Abstract

SHROOM-Visions 2026 asks systems to mark the exact characters hallucinated in a vision-language model’s response. The generating model is withheld, so any model-derived signal is a *proxy*, and we ask whether it improves character-level localization. We first characterize the metric: span *position* dominates attainable Cor under the benchmark metric, and constant scores within a response give no ranking signal; an oracle sentence-level detector reaches Cor 0.476 where gold spans reach 0.970. Testing three proxy signals and one direct adjudication baseline, sentence-level signals add at most 0.002 Cor, as predicted. Representation shift is not explained by that account: a probe recovers a fine-grained signal across all four languages (macro AUROC 0.652) yet adds only +0.009. Its discrimination is uneven across error categories, strongest on smaller ones such as OCR errors, but reweighting that mixture accounts for at most 0.028 AUROC. The tested factors do not explain the remaining gap.

1 Introduction

SHROOM-Visions 2026 (Vázquez et al., 2026) asks participants to identify hallucinated *character spans* in responses produced by large vision-language models (LVLMs) across English, French, Italian and Chinese: given an image, a prompt and a response, a system must output per-character hallucination probabilities and a category label from a five-way taxonomy (Mickus et al., 2026). Table 1 summarizes our systems and diagnostics.

Two properties shape what is possible. First, the generating LVLM is not released — unlike the text-only editions (Mickus et al., 2024; Vazquez et al., 2025), the second of which released its generator’s logits — so any model-derived signal is necessarily a *proxy*. Second, the evaluation is a rank correlation over characters *within* each response, reward-

	Cor by language				averaged		
	EN	FR	IT	ZH	Cor	IoU	C+L
Empty [†]	.253	.255	.263	.365	.284	—	—
XLM-R base	.315	.346	.407	.411	.370	.312	.304
XLM-R large	.343	.389	.400	.386	.380	.324	.312
+ repr shift [‡]	.346	.389	.413	.406	.389	.332	.320
+ caption NLI	.338	.389	.421	.415	.391	.334	.318
+ BIO output	.336	.394	.416	.422	.392	.335	.309
Ceiling [†]	.928	.925	.938	.948	.935	—	—

Table 1: Localization results and diagnostic base-lines: test-set Cor per language (ranking is language-independent), and all three official metrics averaged over languages. [†] rows are held-out development diagnostics (not submissions); [‡] marks our final submission. Bold is the best submitted system per column: Cor and IoU peak at BIO output, label-aware correlation at representation shift, which is the disagreement analysed in §6.

ing discrimination between the characters of one response rather than simply identifying whether a response contains any hallucination.

Our goal is diagnostic rather than competitive: we use a supervised token-level tagger as a fixed testbed and a controlled sequence of proxy visual signals to study when multimodal evidence fails to translate into benchmark-level localization (Table 1). We first characterize what the metric rewards (§4), then test these approaches against those constraints (§5). Sentence-level signals contain information but provide little localization benefit, as predicted; representation shift does not fit that account, revealing a different source of limitation.

2 Related Work

SHROOM systems. The strongest text-only systems converged on a pipeline of evidence retrieval, factuality judgement with LLMs, and span alignment (Huang et al., 2025; Hikal et al., 2025; Liu and Chen, 2025). Retrieval was effective because

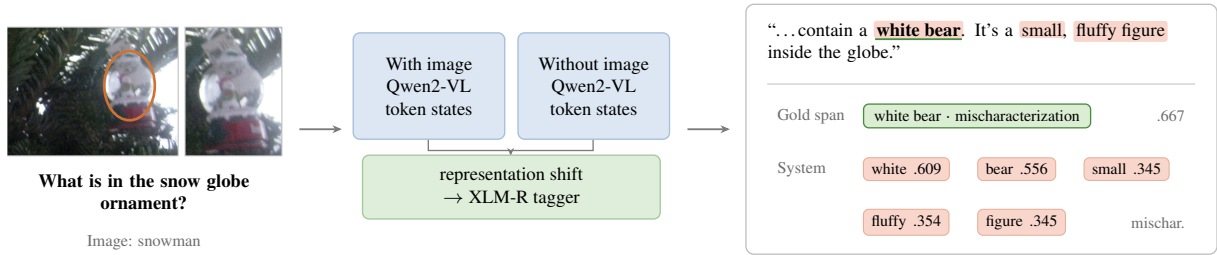


Figure 1: The pipeline on one English training response (train-en-557). A proxy LVLM reads the response twice, with and without the image; the shift between its token representations is concatenated to the XLM-R tagger’s own states (§3). Annotators mark two words, *white bear*, as a mischaracterization of the snowman inside the globe. Our tagger finds the right region but spreads probability across five tokens, and the smallest clause containing the error (55 characters) is five times longer than the annotation (10) — character-level scoring rewards the annotation, not the region. The gold label was selected by 2/3 annotators.

evidence was absent from the input; here it is present but in another modality, making the corresponding challenge visual grounding. Even with the generator’s logits released, Bala et al. (2025) report IoU 0.032.

Model-agnostic detection for LVLMs. Existing LVLM hallucination detectors typically exploit the generator’s own dynamics: cross-modal attention (Zhang et al., 2024), uncertainty (Li et al., 2024), token-level detectors and localization (Park et al., 2025; Kim et al., 2024), and internal-state probes that reach up to 0.93 AUROC for response-level hallucination risk (Su et al., 2024; Kogilathota et al., 2026; Zhang et al., 2026) — a coarser granularity than the one evaluated here; the organizers’ baseline (Nath et al., 2025) measures shift across transformer layers, where ours measures it across visual conditions. Proxy-model representations (Singh et al., 2026) motivate our representation-shift analysis. Prior visual contrastive approaches (Leng et al., 2023; Yin et al., 2025) target generation mitigation rather than character-level localization; we instead evaluate whether visual signals provide usable evidence for span detection.

3 System

Tagger. Our system token-tags XLM-RoBERTa-large (Conneau et al., 2020), with the response as the first segment and the prompt as the second (the prompt is seen but not scored).¹ Two linear heads sit on the final token representations: a span head trained with binary cross-entropy against the largest gold character probability inside each token, and a

five-way category head trained with cross-entropy against the token’s majority category, combined as $\mathcal{L}_{\text{span}} + 0.5 \mathcal{L}_{\text{cat}}$. Token scores are mapped back to characters using the tokenizer’s offset mapping; contiguous above-threshold characters merge into spans, split where the predicted category changes. We train one multilingual model jointly on all four languages, with thresholds tuned per language on a stratified 20% split by a seventeen-value sweep (0.10 to 0.90, step 0.05) maximizing Cor. Character–token conversion is lossy: projecting gold spans through the tokenizer and back caps attainable correlation at 0.928–0.948 (Table 1), as annotations may begin mid-token. Responses are additionally truncated at 256 subwords, so characters beyond that point are never scored. Appendix A gives the full configuration.

Integrating the proxy signals. Each proxy signal enters the tagger the same way, as extra per-token features, so that the rows of Table 1 differ only in which features are present. The features are indexed by the *proxy* model’s tokens, which do not align with XLM-R’s, so we route them through characters — the common index, and also what the scorer measures: each proxy token’s vector is scattered onto the characters it covers, and each XLM-R token takes the mean over the characters in its own span. The resulting vectors are standardized with training statistics (the raw scalars sit on very different scales), projected to 64 dimensions by a Linear–GELU–LayerNorm block, and *concatenated to the final-layer* token representations before both heads; a gate zeroes the block for responses whose features could not be extracted. Fusion is therefore late and additive: the encoder never conditions on the visual features, it only re-

¹Code, the notebooks that produced each row of Table 1, and the probe analyses: <https://github.com/rmaacario/shroom-visions-2026>

Spans	Prob	Labels	Cor	C+L
none	—	—	.258	.258
all	flat	—	.004	.002
oracle	flat	—	.262	.260
sentences	flat	gold	.476	.458
gold	flat	wrong	.970	.262
gold	flat	gold	.970	.984
gold	graded	gold	.988	1.00

Table 2: Oracle substitution on English held-out training data, scored with the organizers’ program. *none* predicts nothing; *all* marks the whole response; *oracle* marks the whole response only when it contains a hallucination; *sentences* marks exactly the sentences containing an annotated character; *wrong* replaces each label with a different category. Labels are worth nothing to Cor and 0.72 to Cor+Lbl; graded probabilities add 0.018; gold span boundaries account for nearly all attainable Cor.

ceives them alongside its own output.

Why a text-only testbed. The tagger itself never sees the image. This is deliberate — it makes the marginal contribution of each proxy signal measurable against a constant baseline, which a jointly trained vision–language tagger would confound with the visual capacity of the tagger — but it does bound what we can conclude: our results describe what a proxy signal adds to a tagger that must otherwise infer hallucination from text alone (see Limitations).

4 What the metric rewards

Before adding any signal we asked what the evaluation can reward, by oracle substitution: predictions carrying gold information in one component but not the others, scored with the organizers’ released program (Table 2). Two of the three properties below follow from that program’s definition rather than from our data, and can be checked against it.

Three properties follow. First, Cor ignores category labels: the scorer computes it from span offsets and probabilities, never reading the label field, which affects only Cor+Lbl. Gold spans with *every* label replaced score 0.970, exactly as gold spans with the right labels. Second, graded probabilities are worth 0.018; a flat value over the correct characters recovers almost everything. Span *position* accounts for nearly all attainable Cor.

Third: within a response containing a hallucination, a constant prediction receives score 0 — Spearman’s ρ is undefined without variation, and the scorer returns 0 there by construction. An ora-

cle clean/hallucinated decision followed by marking the whole response scores 0.260 against 0.253 for predicting nothing — marking whole units is barely better than abstaining.

This is testable, and we test it directly. Within a sentence, a sentence-level signal assigns identical values to annotated and non-annotated characters, so it can only exploit contrast between sentences. Marking exactly the sentences that contain annotated characters — an oracle sentence-level detector with perfect sentence identification — scores Cor 0.476 against 0.970 for gold spans (Table 2). Even perfect sentence-level information therefore remains far below character-level localization. Feeding our realised caption-entailment scores through the same scorer gives Cor 0.028 against a 0.258 floor. The next section tests these signals against this constraint.

Figure 1 shows what this means on one response. The median annotated span is only 2.1% of response characters in English and 3.0–3.5% in the other three languages, which makes character precision the dominant difficulty.

5 Supplying visual information

Probe scores are AUROC for separating annotated from unannotated tokens using a logistic classifier on held-out data, with splits **by response** (splitting by token inflates AUROC by ≈ 0.04). Results are means over five random splits, with standard deviations.

5.1 Output-level likelihood

Each response is scored twice under Qwen2-VL-2B-Instruct (Wang et al., 2024) with teacher forcing, once with and once without the image, producing per-token likelihood, entropy, and contrastive differences. We remove the image rather than perturbing it, since noise makes the model overlook regions stochastically (Yin et al., 2025). All output-level features remain near chance (AUROC within 0.04 of random, the furthest being entropy shift at 0.465 and the contrastive log-probability signal at 0.485), so we exclude them from further analysis.

5.2 Internal representation shift

We repeat both passes with the same 2B model, recording internal states (not outputs): at one quarter, one half and three quarters of its hidden-state stack, cosine similarity and Euclidean distance between with/without-image token representations, plus a fixed 32-dim random projection of

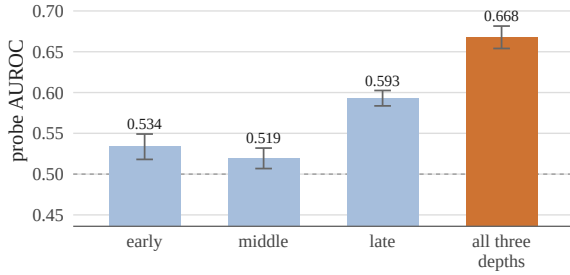


Figure 2: Representation-shift probe by depth (English). No single depth is a strong detector; combining all three exceeds the best by 0.075 AUROC. Bars: means over five response-level splits $\pm$ SD. Values recomputed from released features (§5.2).

the middle-layer difference — 38 features per token, listed in Appendix A (Table 6).

Hidden-state features outperform output-level features across languages: AUROC 0.682 ± 0.016 (en), 0.657 ± 0.008 (zh), 0.639 ± 0.013 (fr), 0.629 ± 0.014 (it), macro **0.652**. No single depth is strong (early 0.534, middle 0.519, late 0.593), but combined they reach 0.668 ± 0.014 — signal distributed across layers (Figure 2). The probe measures token discrimination; Cor rewards character ranking within a response, so high AUROC need not transfer.

Granularity does not account for the limited gain. §4 predicts that sentence-level signals have limited opportunity to improve Cor, so we checked whether granularity is what constrains this one. Aggregating the same held-out predictions to coarser units reduces AUROC from 0.652 per token to 0.597 per sentence and 0.604 per response, macro-averaged. The drop is substantial in English ($0.682 \rightarrow 0.565$) and French but absent in Chinese, where token and sentence aggregation are identical (0.657). Thus even where aggregation preserves the signal, it did not yield larger downstream gains in our integration setting.

Discrimination varies by category, but mixture explains little. Table 4 reports per-category AUROC across all four languages, ordered by annotated-token share. Discrimination is strongest for OCR errors ($.818 \pm .036$) and *other* ($.849 \pm .039$), weaker for invention ($.638 \pm .012$) and mischaracterization ($.622 \pm .016$) — together 75% of annotated tokens — and unstable for miscounting ($.671 \pm .050$, $.556 \pm .017$ in French). Since AUROC pairs positives with a fixed negative pool, reweighting category frequencies yields an exact

counterfactual, not a simulation. Uniform reweighting raises macro AUROC only from 0.652 to 0.680 (and lowers it in French), so category imbalance explains only a small portion of the gap.

What the signal is not. By token type, the probe scores $0.658 \pm .013$ on content tokens and $0.684 \pm .016$ on everything else — punctuation, subword fragments, short function words — against 0.682 over all tokens.² If it measured visual grounding of concepts we would expect the opposite ordering, so we describe it as predictive of annotation rather than as a measure of grounding. Restricting the probe to content tokens therefore does not offer a route to the missing downstream gain: it discards 94% of tokens and lowers AUROC, so a content-only variant would begin from a weaker signal, not a stronger one.

5.3 Caption entailment

Each of the 2,476 distinct images is captioned once by Qwen2.5-VL-7B-Instruct (Bai et al., 2025), conditioned on the questions asked about it (Appendix D); fine-grained attribute questions are 42% of the dataset, on which a generic caption is silent.

Each response sentence is scored against every caption chunk with multilingual NLI (Conneau et al., 2018) and cosine similarity (Reimers and Gurevych, 2019), keeping the maximum entailment and contradiction. On English (21,489 sentences, 34.5% annotated), caption-derived features reach AUROC 0.605 against 0.533 for three surface features (English only; see Limitations). Caption-derived features contain measurable sentence-level information.

5.4 Zero-shot LVLM adjudication

Finally, inspired by winning text-only systems, we tested direct adjudication: an instructed LVLM sees the image, question, and response, and labels unsupported spans.³ On 100 English training examples, it scores Cor 0.202 versus a 0.410 floor — worse than an empty submission.

The failure is diagnostic, not total: quoting was reliable and the model located the right region but marked whole clauses where annotators mark single words — a calibration failure against the guidelines rather than evidence of missing perception, and consistent with the winning text-only systems

²English, five response-level splits, same protocol as above; a content token is an alphabetic run of four or more characters beginning a word, 5.6% of tokens.

³Accessed through a hosted inference API.

needing prompt optimization and fuzzy span alignment atop a strong adjudicator (Huang et al., 2025; Hikal et al., 2025); see Appendix B.

6 Results and analysis

Averaged over languages (Table 1), representation shift contributes +0.009 Cor and caption entailment +0.002. The base-to-large step adds +0.010, but that run also changed the learning rate, batch size and epoch count (Appendix A), so we do not attribute it to encoder size. On the primary Cor+Lbl ranking our system places 16/28 (EN), 11/21 (FR), 12/21 (IT) and 11/22 (ZH); Table 5 gives our rank under all three metrics. In English all four visual signals together move us one place, while 0.21 Cor and seventeen teams separate us from first.

The sentence-level prediction is supported.

Caption entailment carries measurable sentence-level information (AUROC 0.605 against a 0.533 surface baseline) but provides negligible Cor improvement. The predicted loss is also measurable directly rather than only inferred: scoring the realised caption-entailment values through the organizers’ program gives Cor 0.028 against a 0.258 floor (§4), the outcome the decomposition anticipates for a signal that is constant within a sentence. Adjudication similarly identifies relevant regions but fails at span calibration. Both illustrate the same mismatch: informative visual evidence does not necessarily translate into character-level localization. The sentence-level ceiling is not zero (0.476, §4), so granularity imposes an upper constraint independent of signal quality.

Representation shift differs, and we found no evidence that either tested explanation accounts for it. It contains fine-grained information, yet aggregation hurts in three languages and is neutral in Chinese, and a uniform category mixture would raise AUROC by only 0.028 (§5.2) — so neither granularity nor category mixture accounts for the limited downstream improvement.

The three official metrics disagree. Cor and IoU both peak at the BIO configuration (.392, .335 averaged over languages), while label-aware correlation peaks at representation shift (.320) and falls to .309 under BIO — which is why we submitted the former as our final system (Table 1). This follows from Table 2: categories are worth nothing to Cor and 0.72 to Cor+Lbl, so the two reward different halves of the same prediction. Reporting one

alone would have made the same runs look monotonically better than they are. The disagreement is not confined to our submissions: in the official English rankings one team places 21st on Cor+Lbl but 10th on Cor and 8th on IoU, and in Italian the IoU winner is not the Cor winner.

Restructuring the output does not change the picture.

We also replaced the span head with a three-way BIO classifier, since span-initial tokens are known to be more detectable than continuations (Snel and Oh, 2025). On development data this was our largest single improvement (+0.016 Cor, positive in all four languages); on test it was +0.001, and it cost 0.009 on Cor+Lbl — a further instance of the same pattern rather than a result.

7 Conclusion

Span position dominates attainable Cor, and even an oracle sentence-level detector reaches only Cor 0.476 against 0.970 for gold character spans — as caption entailment and adjudication both show. Representation shift is not accounted for this way: it contains fine-grained information, but neither granularity nor category mixture explains why that information yields only modest gains — reweighting the latter accounts for at most 0.028 AUROC. Proxy visual signals should therefore be evaluated not only by how informative they are, but by whether their information is aligned with the granularity and error distribution of the target benchmark — a distinction that is not specific to SHROOM and may arise in benchmarks scoring character- or token-level grounding.

Limitations

Uncertainty and explanation status. Every difference in Table 1 comes from one training run, and the official test annotations are withheld, so paired bootstrap or permutation tests on the leaderboard scores are impossible; per-language deltas below roughly 0.03 should not be over-interpreted, and the +0.009 we attribute to representation shift is within that range. Development and test correlation diverge in both directions: across the two runs whose development scores we retain, the gap ranges from −0.007 (French) to +0.054 (English), so development scores are not a reliable guide to test scores here, and development thresholds were in any case selected as a maximum over a seventeen-value sweep. The two claims we do rest

weight on are not of this kind: the metric decomposition in §4 is deterministic given the scorer, and the probe AUROCs carry standard deviations over five response-level splits. Even so, the two explanations differ in status. For the sentence-level signals the decomposition predicts something about Cor while the probes are scored by AUROC, and we do not convert a sentence-level score into a submission and measure the predicted loss directly, so that account remains consistent rather than demonstrated. For representation shift we did run the counterfactual and it argues against our own hypothesis; what remains untested there are threshold effects, feature redundancy with the tagger, and optimization failure, any of which could explain the small downstream gain. Restricting the features to content tokens we did test on the probe, where it lowers AUROC (§5.2); whether it changes the downstream Cor gain would still require retraining the tagger. Each of these requires retraining the tagger and rescored on the official test set, which has closed, so we report them as untested rather than as ruled out.

Proxy models, scales and configurations. We deliberately held the proxy family fixed so that differences between signals reflect signal type rather than architecture; this prevents separating our findings from properties of that family, and since responses were produced by five undisclosed LVLMs the proxy may share failure modes with the generators. Output-level and representation-shift experiments used a 2B model while captioning and adjudication used larger ones, so comparisons across those settings are not controlled. Each signal was tested in one configuration: caption entailment with one caption per image, one captioner, one NLI model and max-pooling over chunks; adjudication with one model, one prompt and 100 examples. The fusion was likewise a single configuration: one projection width, one fusion point (the final layer), one gating rule, none of them swept. Against the seventeen values we tried for the decision threshold, the integration received no comparable search, so the $+0.009$ bounds *this* integration of the signal rather than the signal itself. Our negative results bound these configurations, not sentence-level signals in general. The testbed is also text-only by construction (§3): conclusions about what raises Cor are conclusions about a tagger that infers hallucination from text alone, and a tagger with its own visual encoder might exploit the same signals

differently. Representation shift also scores higher on non-content tokens than on content words, so we report it as predictive of annotation rather than as a measure of grounding, and any causal reading should be made cautiously.

Annotation and language coverage. A token covered by spans of different categories is assigned the label of the highest-probability span; earlier analyses estimated category-specific values from a single split under a different convention, which placed miscounting at .758, so we report five-seed averages throughout ($.671 \pm .050$ for that category). Aggregate AUROC is unchanged, being independent of category assignment. Caption entailment is reported on English alone because pooling languages inflates the surface baseline from 0.533 to 0.623: sentence length and hallucination rate both covary with language, so length partially encodes language identity. Segmentation is punctuation-based and less reliable for Chinese than for the three Latin-script languages.

References

- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. [Qwen2.5-vl technical report](#). *Preprint*, arXiv:2502.13923.
- Maryam Bala, Amina Abubakar, Abdulhamid Abubakar, Abdulkadir Bichi, Hafsa Ahmad, Sani Abdullahi Sani, Idris Abdulmumin, Shamsuddeen Hassan Muhammad, and Ibrahim Said Ahmad. 2025. [Hau-saNLP at SemEval-2025 task 3: Towards a fine-grained model-aware hallucination detection](#). In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1737–1741, Vienna, Austria. Association for Computational Linguistics.
- 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.
- Alexis Conneau, Rutu Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. [XNLI: Evaluating cross-lingual sentence representations](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brus-

- sels, Belgium. Association for Computational Linguistics.
- Baraa Hikal, Ahmed Nasreldin, and Ali Hamdi. 2025. [MSA at SemEval-2025 task 3: High quality weak labeling and LLM ensemble verification for multilingual hallucination detection](#). In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 989–995, Vienna, Austria. Association for Computational Linguistics.
- Sicong Huang, Jincheng He, Shiyuan Huang, Karthik Raja Anandan, Arkajyoti Chakraborty, and Ian Lane. 2025. [UCSC at SemEval-2025 task 3: Context, models and prompt optimization for automated hallucination detection in LLM output](#). In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1981–1992, Vienna, Austria. Association for Computational Linguistics.
- Minchan Kim, Minyeong Kim, Junik Bae, Suhwan Choi, Sungkyung Kim, and Buru Chang. 2024. [ESREAL: Exploiting semantic reconstruction to mitigate hallucinations in vision-language models](#). In *Proceedings of the European Conference on Computer Vision (ECCV)*.
- Sai Akhil Kogilathota, Sripadha Vallabha E G, Luzhe Sun, and Jiawei Zhou. 2026. [Halp: Detecting hallucinations in vision-language models without generating a single token](#). *Preprint*, arXiv:2603.05465.
- Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2023. [Mitigating object hallucinations in large vision-language models through visual contrastive decoding](#). *Preprint*, arXiv:2311.16922.
- Qing Li, Jiahui Geng, Chenyang Lyu, Derui Zhu, Maxim Panov, and Fakhri Karray. 2024. [Reference-free hallucination detection for large vision-language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4542–4551, Miami, Florida, USA. Association for Computational Linguistics.
- Xu Liu and Guanyi Chen. 2025. [CCNU at SemEval-2025 task 3: Leveraging internal and external knowledge of large language models for multilingual hallucination annotation](#). In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 448–454, Vienna, Austria. 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.
- Timothee Mickus, Elaine Zosa, Raul Vazquez, 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.
- Sujoy Nath, Arkaprabha Basu, Sharanya Dasgupta, and Swagatam Das. 2025. [Hallushift++: Bridging language and vision through internal representation shifts for hierarchical hallucinations in mllms](#). *Preprint*, arXiv:2512.07687.
- Eunkyu Park, Minyeong Kim, and Gunhee Kim. 2025. [Halloc: Token-level localization of hallucinations for vision language models](#). *Preprint*, arXiv:2506.10286.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Akshita Singh, Prabesh Paudel, and Siddhartha Roy. 2026. [Hallucination detection via activations of open-weight proxy analyzers](#). *Preprint*, arXiv:2605.07209.
- Jakob Snel and Seong Joon Oh. 2025. [First hallucination tokens are different from conditional ones](#). *Preprint*, arXiv:2507.20836.
- Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. 2024. [Unsupervised real-time hallucination detection based on the internal states of large language models](#). *Preprint*, arXiv:2403.06448.
- 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 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.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, et al. 2024. [Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution](#). *Preprint*, arXiv:2409.12191.

Hao Yin, Guangzong Si, and Zilei Wang. 2025. [The mirage of performance gains: Why contrastive decoding fails to mitigate object hallucinations in mllms?](#) *Preprint*, arXiv:2504.10020.

Feiran Zhang, Yixin Wu, Zhenghua Wang, Xiaohua Wang, Changze Lv, Xuanjing Huang, and Xiaoqing Zheng. 2026. [Vib-probe: Detecting and mitigating hallucinations in vision-language models via variational information bottleneck](#). *Preprint*, arXiv:2601.05547.

Yudong Zhang, Ruobing Xie, Xingwu Sun, Yiqing Huang, Jiansheng Chen, Zhanhui Kang, Di Wang, and Yu Wang. 2024. [Dhcp: Detecting hallucinations by cross-modal attention pattern in large vision-language models](#). *Preprint*, arXiv:2411.18659.

A Implementation details

Table 3 lists the tagger’s configuration. It is *not* constant across Table 1: the base row was trained at 2×10^{-5} with batch 16 for three epochs, and every large row at 1×10^{-5} with batch 8 for five. The base-to-large difference of +0.010 Cor therefore confounds encoder size with the optimizer and schedule, and we do not read it as an effect of scale. The four large rows do share their configuration, so the deltas *among* them — the ones the paper rests on — are not confounded in this way.

encoder	xlm-roberta-large (-base for the first row)
max length	256 subwords; the response segment is truncated, the prompt is not
heads	Linear($h, 1$) span, Linear($h, 5$) category, dropout 0.1
loss	$\mathcal{L}_{\text{span}}$ (BCE) + 0.5 $\mathcal{L}_{\text{cat}}$ (CE)
optimizer	AdamW, weight decay 0.01; lr 2×10^{-5} (base), 1×10^{-5} (large)
schedule	linear decay, 10% warmup, grad clip 1.0
batch / epochs	16 / 3 (base), 8 / 5 (large)
frozen	embedding matrix, in the fused and BIO rows only
dev split	20%, stratified by clean/hallucinated
threshold	per language, 0.10–0.90 step 0.05, max Cor
feature block	38 → 64 (representation shift), 45 → 64 once caption features are added, Linear-GELU-LayerNorm
seed	13 (single run; see Limitations)

Table 3: Tagger configuration. The threshold is the only quantity we searched.

The 38 features. Per proxy token: cosine similarity and Euclidean distance between the with- and without-image representations at one quarter, one half and three quarters of the hidden-state stack (six scalars), and a 32-dim Gaussian random projection of the middle-layer difference, drawn once with a fixed seed and reused for every example and language. Both passes teacher-force the gold response, so the features are defined for every response token.

Alignment. The proxy and the tagger tokenize differently, so features pass through characters in both directions: a proxy token’s vector is written to the characters its offset span covers, and an XLM-R token reads the mean over the characters in its own span. Features are standardized with training-set means and standard deviations before the projection, since the cosines (near 1) and the distances (an order of magnitude larger) would otherwise be weighted by scale alone.

What was not tuned. The projection width, the fusion point, the gate, the standardization and the category-loss weight were each fixed a priori and never varied. Caption-entailment features enter the same block in place of the representation-shift features; the BIO variant replaces the span head with a three-way token classifier and leaves the rest unchanged.

B Adjudication failure modes

Of 100 English training examples, 9 returned a quote that failed to match the response and none returned an empty reply. Two failure modes account for the score. First, granularity: given a response annotated only at *hidden* and *ceiling*, the model returned both containing clauses in full, roughly ten times the annotated extent. Second, verifiability: for a response the annotators left entirely clean, it marked a correct statement about the chemical origin of a vegetable’s colour — true, but not checkable against the image, which the guidelines do not treat as a hallucination.

C Additional results

Table 4 reports per-category probe AUROC for all four languages, Table 6 the remaining probe results, and Table 5 our official rank under each metric.

Category	share	EN	FR	IT	ZH
invention	46%	.638	.660	.619	.645
mischaracteriz.	29%	.622	.594	.625	.632
other	18%	.849	.626	.637	.654
OCR	4%	.818	.670	.784	.867
miscounting	3%	.671	.556	.647	.794
all categories		.682	.639	.629	.657
uniform mixture		.720	.621	.662	.718

Table 4: Representation-shift probe AUROC by annotated category, ordered by share of annotated tokens (English shares; means over five response-level splits). Standard deviations are $\leq .017$ for invention and mischaracterization and up to .059 for the other three. The last row reweights each language’s categories to a uniform mixture: the aggregate rises by 0.028 on average, and falls in French.

Our rank	EN	FR	IT	ZH
by Cor+Lbl	16th	11th	12th	11th
by Cor	18th	13th	12th	14th
by IoU	19th	13th	13th	15th
teams ranked	28	21	21	22

Table 5: Where our final system placed on the official leaderboard, under each of the three ranking metrics, against the number of teams ranked in that language. The three metrics place us differently in every language except Italian: in English we are 16th, 18th and 19th depending on which one is used.

Feature subset (English)	AUROC
32-d random projection	.620 $\pm$.018
all 38 features	.682 $\pm$.016
per token	.652
per sentence	.597
per response	.604

Table 6: Remaining probe results. The lower block re-evaluates the same token-level predictions after aggregating them to coarser units.

D Caption prompt

Each image received one caption, generated once and reused across every example referencing it. `{questions}` is the list of distinct questions asked about that image across all four languages, capped at six.

Captioning prompt

Describe this image factually and in detail. State the objects present, their colours, shapes and other visible attributes, how many of each there are, where they are relative to

each other, and transcribe any text visible in the image.

Your description must explicitly answer each of these questions:

`{questions}`

Describe only what you can actually see. If something cannot be determined from the image, say so plainly rather than guessing.