

Dynamos at SHROOM-Visions: Metric-Aligned Token Tagging and Ensembles for Multilingual Vision-Language Hallucination-Span Detection

Muhammad Yamin
University of Girona
u6110888@udg.edu

Sheikh Abdul Wahid
Universidade do Algarve
a91643@ualg.pt

Ali Arshad
Skolkovo Institute of
Science and Technology
Ali.Arshad@skoltech.ru

Abstract

Vision-language models describe images fluently and, too often, wrongly. SHROOM-Visions 2026 asks systems to mark exactly which characters of such a description are hallucinated, and of what kind, in Chinese, English, French, and Italian. Working on this task we find that how a system reads predictions out of a model matters as much as which model it uses: the metric is a rank correlation against soft, multi-annotator labels, and aligning the prediction pipeline to it is worth as much as a larger encoder. We treat the problem as per-character sequence tagging. Starting from a fine-tuned XLM-RoBERTa tagger, we (i) formulate the task with one independent probability head per hallucination class, which matches the labelled and unlabelled correlation metrics; (ii) reconstruct spans in a way that respects the scorer’s tie-breaking behaviour; and (iii) tune per-class decision thresholds with a bagged Bayesian search over resampled validation subsamples. These changes lift the score substantially with no change to the underlying model. Keeping this output layer fixed, we then compare encoders: a text-only tagger with added visual features, and a fine-tuned vision-language model. The vision-language model, adapted with low-rank updates and read out through the same five heads, is our strongest system in every language and improves most on the labelled score, which is the part of the task that needs the image. It also transfers cleanly from validation to test, in contrast to the tuned text ensembles, whose threshold choices overfit the validation split.

1 Introduction

Vision-language models routinely produce fluent text that is not supported by the image they are conditioned on. SHROOM-Visions 2026 (Vázquez et al., 2026) asks participants to locate this behaviour at the level of individual characters, and to assign each hallucinated span one of five classes:

invention, mischaracterization, OCR error, miscounting, and other. The task is scored separately for each of four languages, so a system may be strong in one language and weak in another. This is the fourth edition of the SHROOM series (Mickus et al., 2024; Vázquez et al., 2025) and the first to move into a multimodal, multilingual setting.

Our participation started from a simple observation. The absolute scores on this task are modest for everyone, because the evaluation is a rank correlation against soft, multi-annotator labels rather than a hard accuracy. In that regime, careful handling of the metric matters as much as model capacity. We therefore spent most of our effort on the interface between the model and the scorer, and only then on adding visual information.

The paper makes three practical points that we believe generalise beyond this shared task. First, the task ships two correlation metrics, and a model built to predict independent per-class character probabilities matches both of them exactly, whereas the more obvious single-label formulation does not. Second, the scorer reconstructs a prediction vector by overwriting overlapping spans in list order, so the order in which we emit spans changes the unlabelled score; ordering spans by confidence recovers a per-character maximum and is never worse. Third, the decision thresholds that turn probabilities into spans are themselves a source of overfitting, and treating them by bagging the search over resampled validation subsamples rather than a single held-out split is the difference between validation numbers that transfer to the test set and numbers that do not.

With this output layer fixed, the encoder becomes a clean, swappable choice. We compare a multilingual text encoder, the same encoder with added image features, and a fine-tuned vision-language model. The vision-language model is our best system by a clear margin in all four languages, and unlike the tuned text ensembles it shows almost

no gap between validation and test. We read this as evidence that, once the interface to the metric is right, capacity spent on grounding the text in the image is capacity well spent.

2 Task and Metrics

The task provides a multilingual, multimodal dataset of image-conditioned model outputs annotated for hallucination spans by three annotators (Mickus et al., 2026). Each example consists of an image, a textual prompt, and a candidate response produced by some vision-language model. Systems must predict, for every character of the response, the probability that it belongs to a hallucinated span and the class of that span. Gold labels are soft: three annotators mark spans, so a character carries a probability in $\{\frac{1}{3}, \frac{2}{3}, 1\}$ for each class, and roughly a quarter of characters that are hallucinated in English fall under more than one class at once. About 28% of training responses contain no hallucination at all.

The organisers rank systems on three character-level metrics (Vázquez et al., 2026). Two are Spearman correlations between the predicted per-character probability and the empirical annotator probability: an unlabelled form (COR), which sums probability across all classes, and a labelled form (COR_L), computed per class and averaged. The third is the intersection-over-union (IoU) of hallucinated characters. Our design targets the two correlations, because our per-class probability formulation matches exactly what they compare against; we did *not* optimise for IoU, and, as a consequence, rank lower on it than on the correlations (Section 4). Optimising the correlations rather than overlap shaped every design decision below.

Agreement among the human annotators is itself moderate: the dataset reports span-overlap agreement in roughly the 0.4 to 0.6 range, lower in French than in Chinese (Mickus et al., 2026). Correlations near 0.5 are therefore close to the range a system can plausibly reach, and the cross-language spread partly reflects this agreement gap rather than model quality alone.

3 System

3.1 Per-character formulation

We tag the response with a multilingual encoder and place five independent sigmoid heads on top, one per class. This choice is dictated by the metric. The unlabelled COR sums probability across

classes, and the labelled COR_L treats each class on its own, so a model that emits an independent probability per class per character reproduces the exact quantities the scorer compares against. It also handles the multi-label characters naturally, which a single softmax over classes cannot. The training target for a character is, per class, the summed annotator probability clipped to one; the loss is a masked binary cross-entropy over response characters only, with an inverse-frequency weight per class so that the rare classes (OCR, miscounting, other) are not ignored.

3.2 Encoder and long responses

Our backbone is XLM-RoBERTa large (Conneau et al., 2020), which covers all four languages including Chinese. Responses are long: the 95th percentile is around 1,400 characters in English, well beyond the 512-token limit of the encoder. We therefore split each response into overlapping windows, each prefixed with the prompt, run the encoder on every window, and stitch the per-token predictions back onto characters by averaging over overlaps. Short uncovered gaps, such as the whitespace that a subword tokenizer folds into a neighbouring token, are bridged from their neighbours so that predicted spans stay contiguous.

3.3 Reconstruction aligned to the scorer

Turning per-character probabilities back into spans is where several points are won or lost. We keep the probabilities graded rather than binary, emitting a fresh span whenever the quantised probability changes, which preserves the ranking a Spearman metric rewards; collapsing a region to its mean costs several points of COR. Two further details follow from how the scorer builds its unlabelled prediction vector: it writes each span’s probability into the characters it covers in list order, so a later span overwrites an earlier one. We therefore sort spans by ascending probability, which makes the highest-probability class win each character, a per-character maximum that matches the summed gold signal and cannot lower COR_L or overlap. We also cap the classes emitted per character to the most confident one or two, since a class predicted where the gold has none scores zero in COR_L. Appendix A gives the full procedure.

3.4 Threshold optimisation

The map from probabilities to spans depends on a per-class threshold and the labels-per-character

cap. We select these on held-out data with tree-structured Bayesian search (Akiba et al., 2019), optimising $0.5 \text{ COR} + 0.5 \text{ COR}_L$ over cached probabilities (Appendix B). This step alone, applied to an already trained model, improved Chinese test COR by 0.023 and COR_L by 0.037 (Table 1), comparable to or larger than any single architectural change we made. Tuning on one validation split overfits, so we bag thresholds across resampled subsamples (Section 4).

3.5 Visual grounding

The text-only tagger is blind to the image, so it cannot in principle tell invention from a correct description. We added visual information in two ways, both of which start as an identity map so that they can only add signal on top of the text model. The first injects a single pooled image embedding from a SigLIP encoder (Zhai et al., 2023) through a feature-wise modulation layer (Perez et al., 2018). The second is stronger: it lets each response token attend to a grid of patch embeddings from the same encoder, so the model can look at the specific region a claim refers to. We also fine-tune a large vision-language model, Qwen2.5-VL (Qwen Team, 2025), with low-rank adaptation (Hu et al., 2022), using it as an encoder and tagging its hidden states with the same five heads.

3.6 Ensembling

Because the metric is a rank correlation, averaging calibrated probabilities across models is a reliable way to gain a little more. We train several text taggers with different seeds on a shared validation split, average their per-character probabilities together with the grounded models, and retune thresholds on the averaged predictions.

4 Experiments

4.1 Setup

We hold out 10% of the training data per language for validation and threshold tuning, and keep this split fixed across all ensemble members so that the tuning set is never seen during their training. Models are trained with the HuggingFace Transformers library (Wolf et al., 2020) in mixed precision. We score locally with a reimplement of the official scorer, verified to agree to numerical precision, so we can iterate without the platform. Unless stated otherwise the encoder is XLM-RoBERTa large, trained for six epochs.

4.2 Results

Table 1 reports test-set COR and COR_L for the main systems, and Figure 1 shows the full trajectory. A constant baseline that predicts no hallucination is correct only on the empty responses and scores well below our systems in every language; the exact baseline numbers are in the overview (Vázquez et al., 2026).

The single most productive change was not a model at all. Retuning the thresholds of our first model from an overlap objective to the reported correlation objective, with no retraining, moved test COR up by between 0.004 and 0.023 across languages and COR_L by up to 0.05. Adding visual features to the text encoder helped, but modestly and unevenly. The pooled image vector gave small gains in English and Italian and a small loss in Chinese, consistent with a single global vector being too coarse to localise a specific object or count. Patch-level cross-attention was the better of the two and, once ensembled, gave the best Chinese result among the text-based systems.

The fine-tuned vision-language model is a different story. Read out through the same five heads, it is the strongest system in every language on both metrics, ahead of the best text ensemble by 0.04 to 0.06 in COR and by up to 0.07 in COR_L . The gain is largest on the labelled score, which asks not only where a hallucination is but which kind it is, and which therefore depends most on actually seeing the image. The gap between the pooled and patch text variants and the vision-language model suggests that late, shallow fusion of image features is a poor substitute for a model that attends to the image throughout its depth.

Averaging vision-language models adds a final, consistent margin. Three 7B models from different seeds and a 32B model adapted with QLoRA, averaged at the character level and retuned, form our best system in every language (Table 1, last row), at a mean unlabelled correlation near 0.51. The 32B is no stronger than a 7B alone, but its different scale makes it a useful ensemble member; a same-size higher-resolution variant added nothing and was dropped.

On the official test leaderboard, our best system ranked 3rd–5th across languages on COR_L and 5th–6th on COR, placing among the top systems on both correlation metrics. On the third ranked metric, IoU, which we did not optimise, it ranks lower (6th–9th; test IoU 0.39/0.44/0.42/0.46 for

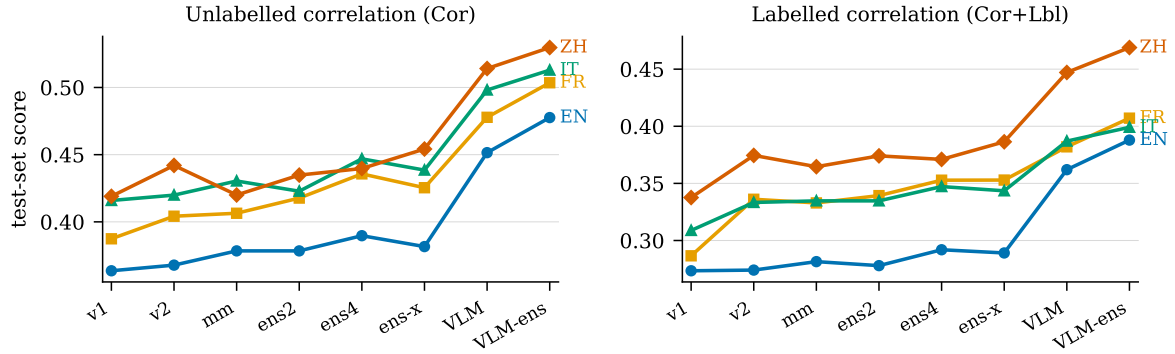


Figure 1: Test-set performance across our submissions, per language. Left, unlabelled correlation; right, labelled correlation. Two jumps stand out: from the first submission (thresholds tuned for overlap) to the second (thresholds tuned for the reported correlation metrics), with no change to the model, and the final step to the fine-tuned vision-language model, which is best in every language.

EN/FR/IT/ZH) – a direct and expected cost of tuning thresholds for the correlations rather than for overlap (Limitations). We are not the top-ranked entry, and a gap to the leaders remains. Naming the hallucination class, rather than only locating it, is the harder half of the task: the rare classes (OCR, miscounting, other) remain the weakest, in line with their scarcity in training and with the invention-versus-mischaracterization confusion that the dataset reports even among human annotators (Mickus et al., 2026).

4.3 Read-out ablation

Table 2 isolates each read-out choice on validation, using the final ensemble’s cached probabilities (mean over the four languages), and is the quantitative backbone of our metric-alignment claim. Tuning thresholds for the correlation metrics rather than for overlap, and the graded reconstruction, give the clear gains; confidence ordering a small one. The labels-per-character cap does not bind for this ensemble, and subsample-bagged tuning is marginally below single-split tuning *on validation* by design – it trades a little validation score for the test-set generalisation below.

4.4 Generalisation from validation to test

Systems that depend on a tuned threshold did not transfer perfectly. Figure 2 plots validation against test COR. The text ensemble sits off the diagonal, and the direction of its gap is not the same in every language, which is the signature of thresholds overfit to a small validation split of only a few hundred examples scored with a noisy correlation. Our mitigation is to run the threshold search on several resampled subsets of the validation data and aver-

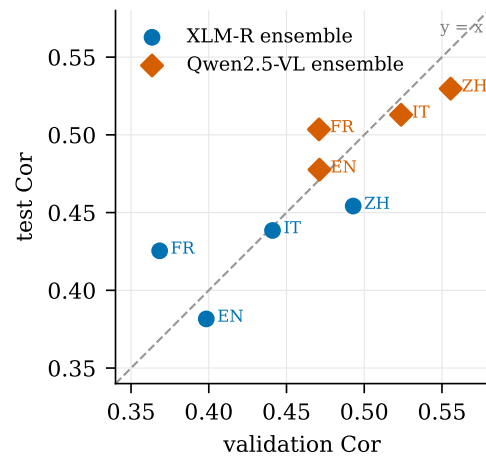


Figure 2: Validation against test COR. Points on the dashed diagonal generalise without loss. The tuned XLM-R ensemble scatters off the line and the gap changes sign across languages; the fine-tuned vision-language ensemble sits close to the line in every language (above it in English and French, marginally below in Italian and Chinese). Test scores are from evaluation-phase submissions, so this is an indicative rather than a fully held-out generalisation check (Limitations).

age the resulting thresholds, taking a majority vote on the labels-per-character cap. This trades a little validation score for cut-offs that are more central and transfer better. The vision-language ensemble, by contrast, sits close to the diagonal in all four languages: its test score exceeds validation in English and French and falls only slightly short in Italian and Chinese, far tighter than the text ensemble. Because most of its predictive signal comes from the fine-tuned encoders rather than from a tuned cut-off, it has far less to overfit, which we consider a practical argument for the approach beyond its raw score.

System	EN		FR		IT		ZH	
	COR	COR _L	COR	COR _L	COR	COR _L	COR	COR _L
XLM-R, overlap-tuned	.364	.274	.387	.287	.416	.309	.419	.338
XLM-R, metric-tuned	.368	.274	.404	.336	.420	.333	.442	.375
+ pooled image (FiLM)	.378	.282	.406	.333	.430	.335	.420	.365
Ensemble (3×text + FiLM)	.390	.292	.436	.353	.447	.347	.440	.371
Ensemble (3×text + x-attn)	.382	.289	.425	.353	.439	.344	.454	.386
Qwen2.5-VL 7B, LoRA	.452	.362	.478	.382	.498	.387	.514	.447
ensemble (3×7B + 32B)	.478	.388	.504	.407	.513	.399	.530	.469

Table 1: Test-set correlation, per language. COR is the unlabelled correlation, COR_L the labelled one. Best per column in bold. All systems share the same metric-aligned output head and reconstruction; the difference is the encoder. The fine-tuned vision-language models are best in every language on both metrics, and by the largest margin on the labelled score, which is where seeing the image matters most; averaging four of them (three 7B seeds and one 32B, adapted with QLoRA) adds a further consistent gain.

Read-out variant	COR	COR _L
Full read-out	0.505	0.439
thresholds tuned for IoU	0.500	0.423
graded → mean-prob spans	0.498	0.435
without confidence ordering	0.504	0.439
without labels-per-char cap	0.505	0.439
single-split thresholds (no bagging)	0.508	0.440

Table 2: Read-out ablation on validation, mean over languages. Each row flips one choice; threshold-dependent rows are retuned. The metric objective (tuning for correlation, not IoU) and the graded reconstruction matter most.

5 Conclusion

Our submission separates two questions that are easy to conflate: how to read predictions out in a way the metric rewards, and which encoder to read them from. A per-class formulation, a reconstruction that respects the scorer’s tie-breaking, and subsample-bagged thresholds account for a large share of our gain at no inference cost. Holding that output layer fixed, a fine-tuned vision-language model is the strongest encoder in every language, by the widest margin on the labelled score, and it generalises to the test set without the gap the tuned text ensembles show. We hope both the metric-alignment recipe and the encoder comparison are useful beyond this dataset.

Limitations

Our absolute scores are modest, which is inherent to a rank-correlation metric on soft labels but also leaves room for stronger multimodal models than the ones we trained. We optimise the two correlation metrics and not IoU, the third ranked metric, and consequently rank lower on IoU (Sec-

tion 4); a threshold objective that included overlap would trade correlation for IoU. The pooled and patch grounding both reuse a frozen vision encoder rather than fine-tuning it end to end. Our threshold tuning, even with subsample bagging, still shows a validation-to-test gap that we mitigate but do not remove. We report a single run per configuration and do not conduct significance testing; only the vision-language ensemble aggregates multiple seeds, so most single-configuration differences (and the small effects reported in the text) should be read as indicative rather than statistically established. The evaluation phase allowed repeated submissions, and our progression (Figure 1) and validation-to-test comparison (Figure 2) draw on those submissions, so the reported test numbers are not fully independent of test-set feedback and the validation-to-test claim should be read with that caveat. Finally, two referenced images could not be read from disk because of illegal filename characters, and we fall back to a zero image representation for those examples.

Ethical Considerations

Our system is a detector, intended to help evaluate and study hallucination in vision-language models rather than to generate content. We use only the data provided by the shared task, which consists of publicly sourced images and model outputs annotated by the organisers; we collect no personal data and add no new annotations. The detector is imperfect and language-dependent, so it should not be treated as a reliable safety filter in deployment: both missed and spurious detections are common on a metric that tops out well below one, and errors may be distributed unevenly across languages. We

fine-tune existing pretrained models and therefore inherit whatever biases they carry. Training used on the order of a few GPU-days on a single accelerator, including the larger quantized model, which we report for transparency and reproducibility.

References

- Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of KDD*.
- 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 ACL*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of ICLR*.
- 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 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 SemEval*.
- Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of AAAI*.
- Qwen Team. 2025. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.
- Raúl Vázquez, 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: MU-SHROOM, 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.
- Raúl Vázquez, Aman Sinha, Chuyuan Li, Artem Shelmanov, Artem Vazhentsev, Claudio Savelli, Eduardo Calò, Emilio Raimond, Stella Frank, Hengyu Luo, Flavio Giobergia, Vincent Segonne, Lorenzo Vaiani, Jörg Tiedemann, and Timothee Mickus. 2026. [Overview of SHROOM-vision 2026: A shared task on hallucination detection in large vision-language models](#). *Preprint*, arXiv:2608.25662.
- Thomas Wolf, Lysandre Debut, Victor Sanh, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of EMNLP: System Demonstrations*.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of ICCV*.

A Reconstruction procedure

Let $p_c[i]$ be the predicted probability that character i belongs to a span of class c , obtained by averaging the per-token outputs of all overlapping windows back onto characters (short runs of characters left uncovered by the tokeniser and bounded on both sides are filled by averaging their two neighbours). We convert this matrix into submission spans as follows.

1. Quantise each probability to q decimal places ($q = 4$).
2. *Labels per character*: at each character, among the classes whose quantised probability meets their class threshold, keep only the k most probable (k tuned per language over $\{1, 2, 3, \infty\}$); suppress the rest.
3. *Graded run-length spans*: for each class independently, scan the characters; open a span where the (surviving) quantised probability first meets the threshold and is positive, and close it whenever the quantised value changes or drops below threshold, emitting one span per maximal equal-value run. This keeps the graded probability rather than a single per-span mean.
4. *Confidence ordering*: sort all emitted spans by ascending probability. The scorer builds its unlabelled prediction vector by writing each span’s probability in list order, so the last write wins; ascending order makes the highest-probability class win each character, a per-character maximum.

The procedure is linear in the response length. Flattening a run to its mean (skipping the graded encoding) costs several points of COR; emitting spans in

class order rather than by confidence lowers COR on multi-class characters.

B Threshold search

The per-class thresholds and the labels-per-character cap are chosen by tree-structured Bayesian optimisation (Optuna TPE, Akiba et al., 2019). The search space is one threshold per class in $[0.02, 0.9]$ and $k \in \{1, 2, 3, \infty\}$; the objective is $0.5 \text{ COR} + 0.5 \text{ COR}_L$ (we do not optimise for IoU, though it is also ranked; Section 4). To avoid overfitting a single validation split, we run 150 trials on each of five random 70% subsamples of the validation set, then average the five threshold vectors and take a majority vote on k . The search runs on cached probabilities, so it is inexpensive and never touches the model.

C Training details

Text encoder. XLM-RoBERTa large (0.55B parameters), sequence length 256 with overlapping windows, batch size 16, learning rate 1×10^{-5} , six epochs, 6% linear warmup, AdamW, mixed precision. Ensemble members differ only in initialisation and data-shuffling seed and share the same held-out split. Image features for the grounded text variants come from SigLIP so400m (0.4B) at 384px, pooled to one vector for the FiLM variant and to a 16×16 grid for the cross-attention variant; the vision encoder is frozen.

Vision-language models. Qwen2.5-VL 7B and 32B. LoRA of rank 16, $\alpha = 32$, dropout 0.05 on the attention (q, k, v, o) and MLP (gate, up, down) projections; the vision tower is frozen and only the LoRA adapters and a linear tagging head are trained (48M trainable parameters for 7B, 142M for 32B). AdamW, learning rate 1×10^{-4} , 3% warmup, bf16. Batch size 2 (7B) or 1 (32B) with gradient accumulation 8/16; three epochs for 7B, two for 32B. Images are processed at up to about 768 visual tokens. The 32B backbone is loaded in 4-bit NF4 (QLoRA) so it fits a single 80 GB GPU. Each run took on the order of a few hours to about a day on one A100-80GB; test-set inference takes under an hour per model.

D Per-class results

Table 3 breaks COR down by class. The rare classes (OCR, miscounting, other) score high mainly because they are absent from most responses, which the model correctly predicts (an all-zero pair scores

1.0); the frequent invention and mischaracterization classes, near 0.6, are where the real difficulty and the remaining headroom lie, consistent with the invention-versus-mischaracterization confusion the dataset reports among human annotators (Mickus et al., 2026).

Class	COR
Invention	0.63
Mischaracterization	0.58
OCR	0.95
Miscounting	0.92
Other	0.92

Table 3: Per-class unlabelled correlation of the final ensemble on validation, mean over languages. High values on the rare classes are driven by correctly predicting their absence.

E Qualitative examples

Miscounting. *Image:* a table holding a tall stack of saucers.
Prompt: How many unused saucers are there on the table?
Response: ... there are two (MISCOUNT) unused saucers on the table.

Mischaracterization. *Image:* a snow-globe Christmas ornament.
Prompt: What is in the snow globe ornament?
Response: The snow globe ornament appears to contain a white bear (MISCHAR.).

OCR. *Image:* a bus with a destination board.
Prompt: What is written on the bus board?
Response: The bus board shows “Plymouth,” “X80,” and “PAISLEY” (OCR).

Figure 3: Gold-annotated examples from the training data, one per hallucination class. The underlined text is the annotated span and the tag is its class; images are described in words.