

Team Bubus at SHROOM-Visions: Probe–Judge Fusion for Hallucination Span Detection

Berk Bubus

Technische Universität Berlin
bubus@campus.tu-berlin.de

Abstract

We describe the Bubus submission to the SHROOM-Visions shared task on character-level hallucination span detection in image-conditioned LVM outputs (English track). Our system fuses two complementary signals: (i) a *cross-model supervised probe ensemble* — per-layer logistic regressions over the hidden states of Qwen3-VL-32B and InternVL3-38B, each teacher-forcing the response under evaluation — and (ii) a $K=8$ *role-diverse multi-sample LLM judge* whose per-character marking fraction matches the annotator-fraction shape of the gold signal. A fixed two-weight late fusion with an abstain gate reaches Spearman correlation 0.411 and character IoU 0.381 on a leakage-controlled held-out split, measured once, post-competition (majority baseline 0.249). On the official test set the system scores Cor 0.413 / IoU 0.379 / Cor_{lbl} 0.282 — ranks 14/12/16 of 28 teams, the last on the primary ranking metric — agreeing with the held-out measurement within 0.002 on Cor and IoU, consistent with a leakage-free protocol. Beyond the positive recipe, we contribute (a) an unusually broad set of *controlled negative results* — zero-shot white-box signals all fail in the shared-task setting, which we attribute to the mismatch between self-generated and teacher-forced *foreign* text; (b) evidence that error-map-driven prompt engineering recovers precisely the targeted deficits at no precision cost; and (c) two external-data attempts that a pre-specified bootstrap gate correctly rejected.

1 Introduction

The SHROOM-Visions shared task (Vázquez et al., 2026), successor to SHROOM (Mickus et al., 2024) and Mu-SHROOM (Vázquez et al., 2025) and built on the human-written SHEEP benchmark (Mickus et al., 2026), asks systems to locate hallucinated *character spans* in the answers of vision–language models to image-grounded

questions, across four languages (we submit English).¹ Two properties of the evaluation dominate system design. First, the gold signal is *soft*: three annotators mark spans independently, so each character carries a fraction in $\{0, \frac{1}{3}, \frac{2}{3}, 1\}$, and the primary ranked metric is per-sample Spearman correlation (Cor) against these fractions, averaged over samples. Spearman’s monotone invariance means only *within-sample ranking* matters. Second, the metric contains hard discontinuities: on a gold-clean sample an *empty* prediction scores 1.0 while any non-empty prediction scores 0.0; roughly 25% of English samples are clean, making abstention the largest single lever. The secondary metric, character IoU, is computed against the *union* of all annotator spans with fractions ignored, so even a $\frac{1}{3}$ -agreement span counts fully — favouring recall when marking. The primary ranking metric, Cor_{lbl}, additionally conditions on the predicted class.

Our submission (Figure 1) is a two-signal system. A supervised white-box probe reads the hidden states of large open VLMs while they teacher-force the response under evaluation, and a black-box multi-sample judge marks spans directly. The two legs err differently — the probe is well-calibrated in *shape* but nearly blind to semantic fabrication (§4); the judge is precise on clean samples but under-recalls — and a two-weight late fusion with an abstain gate beats both. Section 5 reports what did *not* work under controlled conditions, which we consider a primary contribution: nearly every zero-shot white-box signal fails when transplanted from self-generated to teacher-forced foreign text.

¹Code and analysis scripts: <https://github.com/berkbubus/probe-judge-fusion>

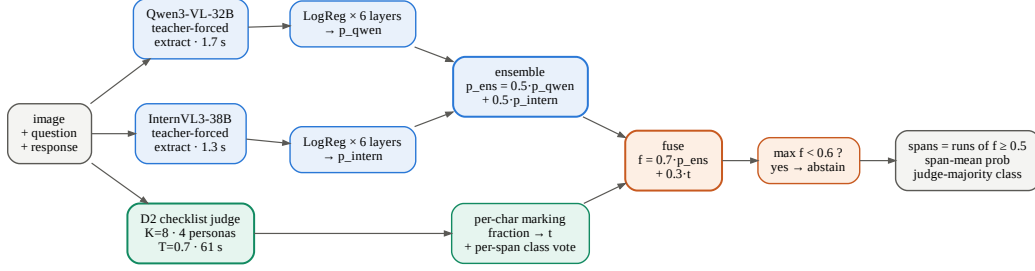


Figure 1: The v2 pipeline. **Blue**: supervised white-box leg (probes trained on gold spans); **teal**: zero-shot black-box judge (prompt in Appendix B); **orange**: fixed fusion and abstain gate. Both backbones read the *foreign* response under teacher forcing; the judge is sampled $K=8$ times under rotating personas and its marking fraction mimics the 3-annotator gold. Per-sample times on one H200; the judge is 96% of runtime.

2 System

2.1 White-box leg: cross-model probe ensemble

Following the supervised-probing lineage of Azaria and Mitchell (2023) and the token-level teacher-forcing setup of HalluShift (Dasgupta et al., 2025), we run the full sample (image + question + response) through a frozen VLM with the response teacher-forced, and harvest each response token’s hidden state at 6 evenly spaced mid-to-late layers (24–64 of 64). Per layer, a logistic regression (class-balanced, $C=1$, standardized features) is trained on gold-aligned token labels; the token probability is the mean over layers, mapped back to characters via the tokenizer offset map. Token labels are binarized from the gold char fractions (Spearman invariance makes calibration beyond ranking unnecessary).

The single novel positive ingredient of our white-box leg is *cross-model ensembling*: we extract with two architecturally distinct backbones, Qwen3-VL-32B (Bai et al., 2025) and InternVL3-38B (Zhu et al., 2025), train an independent probe stack per backbone, and average the two probabilities $p_{\text{ens}} = \frac{1}{2}p_{\text{qwen}} + \frac{1}{2}p_{\text{intern}}$. Their errors decorrelate: the ensemble beats both singles (out-of-fold Cor .412 vs. .400/.384) with a bootstrap win rate of 95.2% over the better single. Adding a third or fourth backbone (Mistral-Small-24B, Idefics3-8B; OOF .368/.344) never helped decisively: across all 24 weighted combinations no candidate exceeded 78% bootstrap against the pair (Appendix C).

2.2 Black-box leg: role-diverse multi-sample judge

The judge is an LLM-as-a-judge span extractor (Zheng et al., 2023) built on Qwen3-VL-32B. It first answers the question *itself* from the image, then diffs the response under evaluation against its own answer (answer-then-compare), verifying every object, attribute, count, text and world-knowledge claim one by one against a checklist, and quoting the exact substrings it judges hallucinated with a class and confidence. Unverifiable specifics are marked *possible* — a deliberate recall push motivated by the union-gold IoU (§4). A prompt-level clean gate outputs nothing when no claim fails (full prompt: Appendix B).

Following the multi-sample strategy of CCNU at Mu-SHROOM (Liu and Chen, 2025) and self-consistency (Wang et al., 2023), we sample the judge $K=8$ times at $T=0.7$ under four rotating expert personas (Appendix A). The per-character probability is the *marking fraction* across the 8 runs — a statistic whose shape directly matches the gold annotator fractions (Huang et al., 2025); a subsampling ablation (Figure 3, Appendix G) shows that the gain from multi-sampling *within the judge leg alone* is modest (+0.013 Cor from $K=1$ to $K=8$, tapering by $K \approx 6$; this is the K -ablation, not the judge’s total contribution to the fusion) — the fraction’s main value is its annotator-like *shape* inside the fusion. Quoted spans map to character offsets by exact substring match. The judge’s majority class vote also provides the span labels of the final output (+.051 Cor_{lbl} on EVAL, main metrics unchanged; Table 1).

2.3 Fusion and decision

The final per-character score is a fixed late fusion $f = 0.7 p_{\text{ens}} + 0.3 t_{\text{judge}}$. If $\max_c f(c) < 0.6$ the system *abstains* (predicts no spans) — the fusion-level analogue of requiring cross-signal agreement, protecting the 25% clean samples worth 1.0 each. It exceeds both legs by a wide margin on the span metrics (Table 1). Otherwise spans are maximal runs of $f \geq 0.5$, each emitted with its span-mean probability and the judge-majority class. All four constants were selected by 5-fold cross-validation on the training side only (§3). Simpler and more complex alternatives (min/max/mean, rank fusion, stackers, MLPs, a soft-rank Spearman surrogate, grid-initialized fine-tuning) never beat this two-weight grid (§5, Appendix D) — the combiner space is saturated at $n=2$ signals.

3 Leakage-controlled evaluation protocol

The protocol: 94 samples whose gold labels had appeared in earlier analysis contexts were removed as contaminated; the remaining data was split by a verified-uniform random draw into EVAL (1,200) and TRAIN-POOL (2,505). All training, model selection, prompt selection and hyperparameter grids ran exclusively on TRAIN-POOL via 5-fold out-of-fold (OOF) prediction; EVAL was touched *once*, to produce the headline fusion number in Table 1, and its train-side optimism gap was ≈ 0.001 . Bootstrap 95% intervals on EVAL are wide relative to recipe deltas (v1 fusion Cor [.397, .439]) — hence *paired* bootstrap win rates, not point comparisons. Post-lock development (backbone additions, prompt variants, external data) used the train-side OOF testbed only, and any change to the locked recipe required a pre-specified adoption gate: $\geq 90\%$ paired-bootstrap win rate over the incumbent on the relevant testbed. “Pre-specified” means the criterion was fixed and recorded in the project’s experiment log *before* each post-lock experiment ran — an internal commitment device, not an external registry. For the submission, probes were refit on all 3,799 labeled samples with the locked recipe.

4 Error-map-driven prompt engineering

Rather than iterating prompts blindly, we first decomposed the OOF errors of the locked system on TRAIN-POOL. The error map: (i) 31% of clean samples received a false span (~ 0.078 mean Cor

System (EVAL, $n=1,200$)	Cor	IoU	Cor _{tbl}
mark-none baseline	0.249	0.249	0.249
probe, Qwen only*	0.399	0.353	0.243
probe, InternVL only*	0.378	0.348	0.235
probe ensemble*	0.402	0.359	0.252
judge only, C2, $K=8^\ddagger$	0.330	0.354	0.296
judge only, D2, $K=8^\ddagger$	0.320	0.352	0.273
v1: $0.8 p_{\text{qwen}} + 0.2 t_{\text{C2}}^*$	0.417	0.378	0.260
v2: $0.7 p_{\text{ens}} + 0.3 t_{\text{D2}}$ (submitted) [†]	0.411	0.381	0.299
v2, official test ($n=1,201$)	0.413	0.379	0.282

Table 1: Main results — every system on the same held-out EVAL bed, official scorer, English. Probe rows at the train-side-selected operating point (gate .8, thr .6); fusion rows at their locked constants. *no class head: Cor_{tbl} under these rows’ actual single-class output; v2’s judge-majority propagation is worth $+0.051$ Cor_{tbl} over single-class (.248 $\rightarrow$.299). [†]adopted via the train-side bootstrap gate; EVAL row is post-competition. [‡]graded spans at grid-best. Final row: official test (ranks 14/12/16 of 28, final leaderboard).

lost); the probe fires *confidently* on clean samples (median max-prob 0.71), so no probability gate can separate them — a sample-level semantic signal is required. (ii) 30% of hallucinated samples were missed entirely; on partially-found samples char recall/precision were .58/.56. (iii) Span *starts* were found precisely (median offset 0) but *ends* stopped early (median -5 chars). (iv) The per-class breakdown (Table 5, Appendix G) revealed the headline deficit: probe char recall is .77/.62 on OCR/miscounting but only .36/.35 on invention and mischaracterization — the two classes holding $\approx 85\%$ of gold characters. Hidden-state probes catch localized token-level errors but are nearly blind to semantic fabrication.

Each deficit was then targeted in the judge prompt: a per-claim checklist against the image, an explicit instruction that gold is a *union* of annotators (“mark unverifiable specifics as *possible*”), full-phrase quoting, per-occurrence re-marking of repeated errors, and a sharpened clean gate. On a 150-sample train slice the winning variant (D2) moved invention recall .38 $\rightarrow$.43 and mischaracterization recall .40 $\rightarrow$.43, cut fully-missed samples 38 $\rightarrow$ 34, raised span coverage 83 $\rightarrow$ 93% — with clean-sample false positives *flat*. Fused scores improved from .405/.379 to .409/.387 (Cor/IoU). The general lesson: error-map-guided prompt engineering converts targeted instructions into precisely the targeted gains. Notably the map’s cheap fix — blanket end-extension — *failed* (§5): the

bias is real but heterogeneous; only content-aware full-phrase quoting repaired it.

5 Controlled negative results

All comparisons used the same folds and official scorer as the positive results (full list: Appendix D).

Zero-shot white-box signals fail on foreign text.

Lookback ratios (Chuang et al., 2024) scored Cor .315; image-ablation (black-image) contrastive grounding features were indistinguishable from mark-none (.255); attention-dispersion and patch-cosine-consistency features scored .235; an unsupervised MIND-style probe (Su et al., 2024) peaked at .350; sampling-based self-consistency in the SelfCheckGPT style (Manakul et al., 2023) collapsed to .080 — far *below* the trivial floor (first three on an early dev slice; rest TRAIN-OOF). Our working hypothesis: these signals are established on *self-generated* text, where the model’s own decoding dynamics carry the hallucination trace. Shared-task responses are *foreign* text teacher-forced through ours, where that trace appears largely absent; on this task and setup, only *supervised* readouts (§2.1) survive — consistent with their being trained under the forced-reading distribution, though our ablations support rather than prove this reading.

Capacity, features, post-hoc. Deeper MLP heads (.350 vs. .400 for logistic regression on identical features), position and repetition covariates, soft-label (fraction-weighted) training, end-shift sweeps ($k \in [-8, +12]$, best $k=0$), smoothing/hysteresis/dilation, word-boundary snapping, and a learned sample-level abstention gate (AUC .789 vs. .783 for max-probability) all failed the adoption gate. The errors are informational, not cosmetic.

Learned fusion. Stackers over 2–8 signals, an n -signal mean, reciprocal-rank fusion, a soft-rank Spearman surrogate loss, and a grid-initialized combiner with real-metric early stopping: none beat the two-weight grid — the space is saturated, not under-optimized.

External data. We mapped M-HalDetect (Gunal et al., 2024) (4,500 samples after class mapping) into our training folds (held-out folds kept pure); OOF degraded (0.781 $\rightarrow$ 0.772 on the fold-level selection metric, bootstrap 18%), which

we attribute to domain shift (InstructBLIP-era caption style and binary labels vs. curated traps and three-annotator union). A second attempt with the task’s own French/Italian/Chinese splits (11,302 samples; identical images and protocol) nudged the metric *positive* (0.781 $\rightarrow$ 0.784) but not decisively (bootstrap 61%) — a clean domain-match gradient reinforcing the domain-shift reading. More labeled data helped surprisingly little even in-domain: the probe’s learning curve is nearly flat past half the data (Figure 3, Appendix G), and growing probe training from 1,500 to 3,399 samples gained only +0.011 honest Cor.

6 Official results

On the official English test set ($n=1,201$; Figure 4) the submitted system scores Cor 0.4130 / IoU 0.3791 / Cor_{lbl} 0.2817, ranking 14/12/16 of 28 teams on the final leaderboard. First, *the held-out protocol and the test score agree*: the submitted system’s EVAL measurement (Table 1) — taken once, *after* the competition, since EVAL was kept single-touch — differs by only +0.002 Cor / −0.002 IoU. This is a post-hoc consistency check rather than a prediction, but it is the outcome a leakage-free protocol should produce, and it supports transferring the ablations above at face value. Second, Cor_{lbl} — the *primary* metric, since the task is to detect *and classify* — is our weakest: our Cor $\rightarrow$ Cor_{lbl} drop is 0.131, versus a median of 0.120 across the thirteen teams above us on Cor; its EVAL $\rightarrow$ test gap (.299 $\rightarrow$.282) is also wider than Cor/IoU’s ± 0.002 — label quality transfers less cleanly. The label comes from the judge’s vote (§2.2; +.051 over single-class, Table 1); its confusion concentrates in mischaracterization voted invention (Appendix G) — a target for a dedicated span-class classifier.

7 Discussion and conclusion

Four lessons generalize. (1) *Supervised white-box probing is the efficiency backbone.* At 3s per sample the probe ensemble delivers +0.153 Cor over baseline (~ 51 mCor/s); the judge’s marginal contribution costs $\sim 300\times$ more per point (Appendix E). It is not blanket superiority (gold labels needed; class labels and most IoU recall still come from the judge), but wherever cost matters, performance-per-second lives in the probe. A practitioner can cut end-to-end cost $\sim 6\times$ by dropping to a single judge pass: probe ensemble +

$K=1$ (~ 11 s/sample) retains Cor .402 / IoU .377 / Cor_{lbl} .290; the remaining seven passes buy only the last $+ .009 / + .004 / + .009$ (Appendix E). (2) *More data is not the lever*. Out-of-domain external data actively hurt ($- .009$); comparing dataset contents explains why (Appendix F). The learning curve is flat past half the data: signal-limited, not data-starved. (3) *Other languages brought no decisive advantage*. Augmenting English training with the task’s own French/Italian/Chinese splits moved the selection metric only $+ .003$ (bootstrap 61%): the cross-lingual hidden-state signature is too weak to matter and English-only training remained the adopted optimum. (4) *Teacher forcing shapes everything*. Responses come from other models, without generation traces; the white-box leg must force-read foreign text — on this task and setup, the constraint most consistent with the pattern of zero-shot internals failing here (§5); a hypothesis our ablations support but do not prove.

Limitations

We submitted only the English track; the recipe is language-agnostic in principle, but we did not verify probe quality per language. The judge leg costs ~ 61 s per sample ($K=8$ on a single H200), which is practical for a shared task but not for interactive use; the probe leg alone (~ 3 s per sample, Cor .402 EVAL) or probe + one judge pass (~ 11 s, recovering most of the IoU and class-label gains) are the deployable points on the cost–quality frontier (Appendix E). The v2 recipe was verified on a 150-sample train slice under a bootstrap gate rather than on the held-out split, a deliberate trade-off to keep EVAL single-touch during the competition; the official test score (0.413, within 0.002 of the post-hoc EVAL measurement) vindicated the choice. All findings are on one dataset family; the negative results about zero-shot white-box transfer are hypotheses supported by controlled ablations, not proofs. Hidden-state extraction requires open-weight VLMs and does not apply to API-only models.

Ethical considerations

All experiments use the organizers’ SHEEP-based shared-task data (Mickus et al., 2026) and openly released VLM checkpoints under their respective licenses; no additional human data was collected. Hallucination span detection is a safety-oriented application: systems like ours are meant to flag

unsupported model claims, not to generate content. Total compute was modest by modern standards (single-GPU jobs throughout, ~ 70 H200-hours end to end including all negative results), and we report it to support cost-aware reproduction.

Acknowledgments

Claude Code (Anthropic) was used to assist with implementation, experiment orchestration, verification of the numerical claims in this paper against the underlying code and data, and L^AT_EX editing. All system design decisions, analyses, and conclusions are the author’s own.

References

- Amos Azaria and Tom Mitchell. 2023. [The internal state of an LLM knows when it’s lying](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, et al. 2025. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*.
- Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James Glass. 2024. [Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1419–1436, Miami, Florida, USA. Association for Computational Linguistics.
- Sharanya Dasgupta, Sujoy Nath, Arkaprabha Basu, Pourya Shamsolmoali, and Swagatam Das. 2025. Hallushift: Measuring distribution shifts towards hallucination detection in LLMs. *arXiv preprint arXiv:2504.09482*.
- Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. [Detecting and preventing hallucinations in large vision language models](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18135–18143.
- 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.

- 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.
- Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. [SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9004–9017, Singapore. Association for Computational Linguistics.
- Timothee Mickus, Claudio Savelli, Eduardo Calò, Emilio Raimond, Stella Frank, Hengyu Luo, Flavio Giobergia, Vincent Segonne, Chuyuan Li, Aman Sinha, Lorenzo Vaiani, Jörg Tiedemann, and Raúl Vázquez. 2026. Can humans dream of electric sheep? human-written samples for fine-grained vision-and-language hallucination benchmarking. *arXiv 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.
- Weihsang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. 2024. [Un-supervised real-time hallucination detection based on the internal states of large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14379–14391, Bangkok, Thailand. 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 Gíbert, 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, 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. *arXiv preprint arXiv:2608.25662*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations (ICLR)*, Kigali, Rwanda.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-bench and chatbot arena](#). In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*.
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, et al. 2025. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Reproducibility details

Data splits. Split seed 20260725; contaminated ids ($n=94$: any sample whose gold labels appeared in an earlier analysis context, including a 12-sample pilot and ~ 20 hand-inspected trap images) removed before splitting. Uniformity of the EVAL draw was verified against 10^4 resamples on clean-rate, class-mix, response-length and generator-model marginals. OOF seed 20260727; identical 5-fold assignments across every compared system.

Probes. Per-layer scikit-learn LogisticRegression ($C=1.0$, $class_weight=balanced$, $max_iter=2000$) on StandardScaler-normalized hidden states; layers {24, 32, 40, 48, 56, 64} of 64 (proportional positions for InternVL); features stored in fp16, trained in fp32; token label = binary overlap with the gold union; ensemble = mean of per-layer probabilities; token $\rightarrow$ char via tokenizer offset mapping (a separate fast tokenizer where the reference implementation loads a slow one). A full fit over $\sim 440k$ tokens $\times$ 6 layers takes 10–22 CPU-minutes.

Judge. Qwen3-VL-32B-Instruct, temperature 0.7, $K=8$; four personas (fact-checker, visual grounding expert, skeptical scientist, meticulous proofreader) each used twice; JSON output parsed with verbatim-quote matching and 1-based occurrence indices; unparseable generations ($<2\%$) count as “no marks” in the fraction. Class label per final span = majority vote over the K runs’ class assignments on overlapping chars.

Fusion grid. $w \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$, $\text{gate} \in \{0.5, 0.6, 0.7, 0.8\}$, $\text{threshold} \in \{0.4, 0.5, 0.6\}$, selected by 5-fold train-side CV; v1 optimum (0.8, 0.7, 0.5) with Qwen-only probe, v2 optimum (0.7, 0.6, 0.5) with the cross-model ensemble — the stronger white-box signal both earns more weight relative to the judge and allows a laxer abstain gate.

Compute. All experiments ran on a single GPU at a time (H200-140GB; scheduling constraints prevented sharding), orchestrated as dependency-chained Slurm lattices with automatic negative-outcome self-cancellation. Feature extraction for the full training set is ~ 2 GPU-hours per backbone; the $K=8$ judge over the 1,201-sample test set is ~ 20 GPU-hours and dominates total cost. All code and analysis scripts are available at <https://github.com/berkbubus/probe-judge-fusion>.

B Full judge prompt (variant D2)

The complete system prompt of the winning judge, reproduced verbatim. The three bolded design elements trace directly to the error map of §4: the claim checklist (missed-span recall), step 2.4 (union-gold recall push), and step 2.5 (clean-sample false positives).

You are an expert annotator for a hallucination detection task. You have the IMAGE itself, and for each sample a QUESTION asked about the image and a RESPONSE produced by a vision-language model.

Procedure - follow in order for EACH sample

STEP 1 - ANSWER IT YOURSELF. Look at the image and write your own careful, correct answer to the QUESTION. Count objects one by one, read visible text character by character, and check first whether the thing the question asks about is present in the image at all.

STEP 2 - CLAIM CHECKLIST. Break the RESPONSE into its individual factual claims and verify EACH one against the image, one by one:

- a) EVERY OBJECT the response mentions - is it actually visible in the image? An object you cannot find = fabricated ("invention").
- b) EVERY ATTRIBUTE - color, material, size, shape, brand, position, spatial relation ("on the left", "behind") - is it correct for the object it describes? Wrong attribute = "mischaracterization".
- c) EVERY NUMBER/QUANTITY - recount yourself. Wrong count = "miscounting".
- d) EVERY quoted or "read" TEXT - if the text exists in the image but is misread = "OCR"; if the response reads text from a sign/label that does NOT exist in the image = "invention".

- e) EVERY WORLD-KNOWLEDGE claim (facts beyond the image): if it is FALSE, mark it as "invention" EVEN THOUGH the image cannot verify it. If it is TRUE, it is NOT a hallucination - leave it unmarked.

Typical hallucinations hide in confident specifics: names, brands, numbers, colors, and details "seen" in low-visibility parts of the image. Be systematic: walk the response start to end so no claim is skipped.

STEP 2.4 - UNVERIFIABLE SPECIFICS. The gold standard is the UNION of 3 annotators - a span flagged by even ONE annotator counts. So for a concrete SPECIFIC detail (a name, brand, number, color, or object) that you can NEITHER confirm NOR rule out from the image, prefer marking it with confidence "possible" over skipping it. Generic true-ish statements are still unmarked.

STEP 2.5 - CLEAN GATE (decide before marking anything). After the checklist, ask: did ANY claim actually FAIL verification on a concrete FACT? If every claim passed - differences being only wording, paraphrase, style, added TRUE detail, hedging/uncertainty, or roughly-correct description - there is NO hallucination: return {"own_answer": "...", "spans": []}. About 1 in 4 responses is fully clean. Marking any span on a clean response scores ZERO for that sample - when all claims pass, return an empty span list.

STEP 3 - MARK SPANS for every claim that FAILED the checklist.

- Quote the COMPLETE PHRASE that carries the wrong content - the number together with its noun ("three seeds"), the full attribute+object phrase ("a red wooden door"), the whole fabricated clause - not just a single head word. Do not stop the quote in the middle of the erroneous phrase.
- If the same wrong fact is REPEATED later in the response (the same wrong number or object mentioned again), add a SEPARATE span for EACH occurrence, with the correct "occurrence" index.
- Near-paraphrases, hedges, refusals, and TRUE extra information are NOT marked.

Classes (one per span)

- "invention" - content about things not present in the image / false world knowledge / text "read" from nonexistent signs
- "mischaracterization" - wrong description of something visible (color, brand, location, shape)
- "OCR" - misreading of text that IS visible in the image
- "miscounting" - wrong quantity of visible countable items
- "other" - disfluencies, garbled tokens, echoed template text

Confidence - how many of 3 independent annotators would flag this span
 "certain" = all 3, "likely" = 2 of 3,
 "possible" = only 1 of 3.

Output - JSON only, one entry per sample id
 {
 "sample id": {

```

"own_answer": "your own short answer",
"spans": [
  {"quote": "exact verbatim substring of the
RESPONSE", "occurrence": 1, "class":
  "miscounting", "confidence": "certain",
  "reason": "one short sentence"}
]
}
}
`quote` must be a VERBATIM substring of that
sample's RESPONSE (including markdown characters
exactly as they appear). `occurrence` is the
1-based index among identical substrings.

```

C Backbone and combination study

Every equal-weight and weighted combination of the four probe backbones, evaluated on identical TRAIN-OOF folds ($n=2,505$) at each combination's own best operating point. "boot" = bootstrap win rate (Cor, 10^4 resamples) against the Q+I incumbent; adoption required $\geq 90\%$.

Combination	Cor	IoU	boot
Q (Qwen3-VL-32B)	.400	.358	0%
I (InternVL3-38B)	.384	.354	0%
M (Mistral-Small-24B)	.368	.340	0%
D (Idefics3-8B)	.344	.316	0%
Q+I (adopted)	.412	.369	—
Q+M	.400	.359	1%
Q+D	.391	.346	0%
I+M	.392	.359	0%
I+D	.387	.353	0%
M+D	.376	.343	0%
Q+I+M	.407	.366	11%
Q+I+D	.413	.368	46%
Q+M+D	.403	.359	1%
I+M+D	.395	.361	0%
Q+I+M+D	.413	.369	53%
2Q+1I	.411	.368	35%
3Q+1I	.408	.365	8%
1Q+2I	.407	.371	28%
2Q+2I+1M	.409	.368	19%
2Q+2I+1D	.414	.369	61%
2Q+2I+1M+1D	.413	.371	72%
3Q+3I+1M+1D	.414	.371	78%
2Q+1I+1M	.407	.365	3%
4Q+2I+1M+1D	.413	.370	58%

Table 2: Cross-model probe combinations (TRAIN-OOF). Several 3–4-model mixtures nominally match or exceed Q+I by $+.001$ – $.002$ Cor, but none decisively (boot $\leq 78\%$); under the pre-specified gate the simpler pair stands. Both weaker backbones would also roughly double extraction cost.

D The full method zoo

Table 3 lists every method and combination evaluated over the project, simplest to most complex,

with its evaluation bed and verdict. WB = white-box, BB = black-box.

E Runtime and the cost–quality frontier

Table 4 gives measured per-sample inference times for every component; the multi-sample judge dominates end-to-end cost. Figure 2 turns this into an explicit frontier by subsampling the stored $K=8$ D2 runs inside the locked v2 fusion. The asymmetry between the legs is stark. The white-box leg buys $+0.153$ Cor over the trivial baseline for 3 s per sample (~ 51 mCor/s); the judge's *marginal* contribution on top of it is $+0.009$ Cor for $+61$ s (~ 0.15 mCor/s) — a $\sim 300\times$ efficiency gap. The judge's value is therefore not raw Cor: (i) its IoU gain arrives almost entirely at $K=1$ (.359 $\rightarrow$.377 of the final .381, at $+7.6$ s) — one judge pass already supplies the recall the probe lacks; (ii) it is the sole source of span *class labels*, and remarkably a single pass also recovers most of that: Cor_{lbl} goes .252 (probe-only, single-class) $\rightarrow$.290 at $K=1$ $\rightarrow$.299 at $K=8$; and (iii) it feeds the abstain gate on clean samples. A practitioner tuning for cost should run the probe ensemble plus a single judge pass (~ 11 s: Cor .402 / IoU .377 / Cor_{lbl} .290); the remaining seven passes purchase the last $+.009/+0.004/+0.009$.

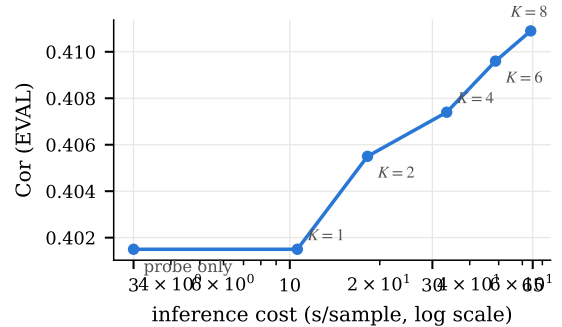


Figure 2: Cost–quality frontier on EVAL: v2 fusion with the D2 judge subsampled to K passes (probe-only = no judge; judge cost 7.6 s per pass). Cor climbs gradually with K , while the IoU (.359 $\rightarrow$.377 $\rightarrow$.381) and Cor_{lbl} (.252 $\rightarrow$.290 $\rightarrow$.299) gains are nearly complete at $K=1$.

F External-data mapping details

M-HalDetect. We converted annotations (ACCURATE / ANALYSIS / UNSURE / INACCURATE segments over InstructBLIP VQA responses) to SHROOM-style char spans: INACCURATE

#	Method	Type	Bed	s/sample	Cor	IoU	Verdict
1	mark-none	baseline	EVAL	0	.249	.249	floor
2	probe, Qwen-only	WB	OOF	1.4	.400	.358	core signal
3	black-image contrast probe (10 feats)	WB	OOF	0.8	.255	.255	$\times$ = floor
4	attention-dispersion/patch-cosine (32 feats)	WB	OOF	0.3	.235	.251	$\times$
5	InternVL3-38B probe	WB	OOF	1.3	.384	.354	alone $\times$; see #6
6	cross-model ens $\frac{1}{2}\mathbf{Q} + \frac{1}{2}\mathbf{I}$	WB	OOF (boot 95.2%)	2.7	.412	.369	✓ first WB win
7	learned combiner (grid-init, real-metric ES)	learned	EVAL-CV	—	.406	.372	$\times$ saturation
8	v1 : $0.8 p_q + 0.2 t, g=.7$	fusion	EVAL	63.3	.417	.378	v1 submission
9	prompt D1: claim checklist ($K=8$)	BB	slice150	55	.405	.382	ok
10	prompt D2 = D1 + recall push	BB	slice150	61	.409	.387	✓ winner
11	class-label propagation (judge vote)	post-hoc	EVAL	ms	± 0	± 0	✓ Cor _{lbl} +.051
12	v2 : $0.7 p_{\text{ens}} + 0.3 t_{\text{D2}}, g=.6$	system	EVAL (gate 97.6%)	65.1	.411	.381	✓ v2 submission

Post-hoc negatives (all $\leq$ base, TRAIN-OOF / slice150): end-shift $k \in [-8, 12]$; smoothing/hysteresis/dilation; word-snap; learned clean-gate (AUC .789 vs .783); confidence-weighted fractions; D1+D2 ensembling; position/repetition covariates; soft-label training; M-HalDetect augmentation ($-.009$, boot 18%); FR/IT/ZH augmentation ($+.003$, boot 61%).

Table 3: Every method evaluated, simplest to most complex, on honest beds (OOF = 5-fold out-of-fold on TRAIN-POOL; slice150 = 150-sample random train subset for bootstrap-gated checks; EVAL-CV = 5-fold CV within EVAL). s/sample = single-H200 inference seconds.

Operation	s/sample	HW
Qwen3-VL hidden extraction (6 layers)	1.3–1.8	H200
InternVL3-38B hidden extraction	1.3	H200
attention/grounding feature variants	0.3–0.8	H200
judge, 1 greedy call	~ 8	H200
judge $K=8$ (prompt C2)	46.6	H200
judge $K=8$ (prompt D2, longer)	~ 61	H200
probe inference + fusion + emit	< 0.1	CPU
v1 total	63.3	
v2 total (judge shared, +1 extract)	~ 65.1	

Table 4: Measured per-sample inference cost. The $K=8$ judge is $\sim 96\%$ of runtime; the cross-model ensemble (v2’s white-box upgrade) adds only $\sim 3\%$ because the judge run is shared between recipes. Probe training is a one-time 10–22 CPU-minute fit per backbone.

segments became gold spans (class invention, fraction 1.0); responses with no INACCURATE segment became clean samples; ANALYSIS/UNSURE segments were left unmarked. We capped the import at 4,500 samples (3,300 hallucinated / 1,200 clean, matching our clean rate) over COCO val2014 images. External samples were added to *training folds only*; every held-out fold remained pure SHROOM, so the OOF comparison isolates the augmentation effect. Result: fold-level selection metric $0.781 \rightarrow 0.772$, bootstrap 18% — rejected.

FR/IT/ZH self-augmentation. The same fold protocol with the shared task’s own non-English training splits (11,302 mapped samples; identical images and annotation protocol). Motivation: both backbones are multilingual, so the hidden-state signature of a hallucinated token may be partially

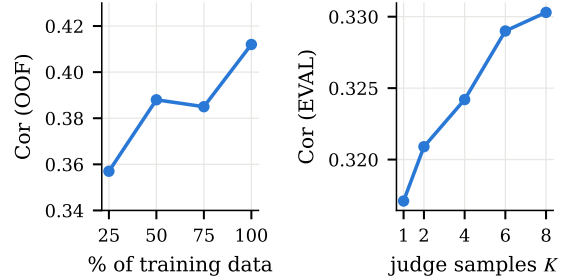


Figure 3: Left: probe-ensemble learning curve (OOF Cor vs. fraction of training data) — the flat tail explains why external data could not help (§5). Right: judge-only multi-sample ablation, computed by subsampling the stored $K=8$ runs at a fixed operating point; gains taper by $K \approx 6$ (§2.2).

language-agnostic, and this is the nearest-domain external data that can exist for this task. Result: fold-level selection metric $0.781 \rightarrow 0.784$, bootstrap 61% — positive in direction but below the $\geq 90\%$ gate; rejected, and the submitted system is unchanged. Together with M-HalDetect ($-.009$, 18%) this traces a monotone gradient in domain match, and suggests the probe is signal-limited rather than data-starved (consistent with the $+.011$ in-domain data lever).

G Additional analyses: figures, class recall, judge-vote confusion

Table 7 crosses every gold span on EVAL with the judge’s majority class vote over the characters it covers (C2 runs, $K=8$). Two structures explain the weak Cor_{lbl} (§6). First, miss rates dominate:

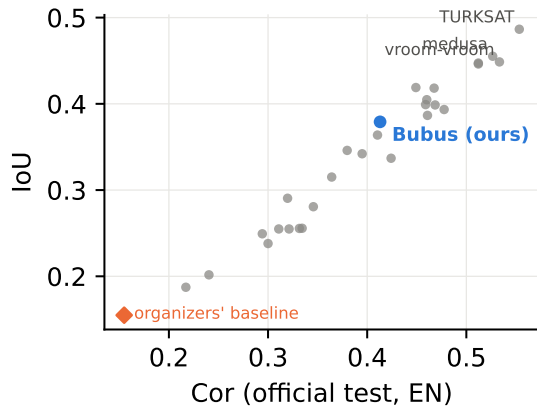


Figure 4: Official English test results, all 28 teams, final leaderboard (§6). Our submission (blue) sits mid-field; the organizers’ baseline is the orange diamond. The post-competition held-out EVAL measurement agreed with our test point to within 0.002.

Gold class	probe char recall	
invention	.36	≈85% of gold chars
mischaracterization	.35	
miscounting	.62	
other	.65	
OCR	.77	

Table 5: Per-class character recall of the probe-only system (TRAIN-OOF, $n=2,505$).

29–63% of gold spans receive no vote at all, and a missed span can never carry a label. Second, among *found* spans the single large off-diagonal is mischaracterization voted invention (415 of 973 found, 43%; the reverse direction is 189) — the judge conflates “describes a visible thing wrongly” with “describes a thing that is not there,” a distinction that often hinges on whether the entity is in the image at all. OCR and miscounting, which have crisp lexical signatures, are labeled reliably once found. other (disfluencies) is essentially never recognized.

H What the fusion fixed: v2 error map and qualitative examples

Table 6 repeats the §4 error decomposition with the full v2 system. The judge leg repaired precisely its targets: fabrication-class recall rose +16/+12 points, fully-missed samples fell by 5 points, and partial char recall jumped .58→.72 — at a moderate precision cost (.56→.50), the expected price of the union-gold recall push. The residual weaknesses are the clean-sample false-

positive rate (~30%, barely moved: when *both* legs are confidently wrong the abstain gate cannot help) and mischaracterization recall, still below .5.

	probe-only (TRAIN-OOF)	v2 fused (EVAL)
clean samples w/ false span	31%	30%
hallucinated fully missed	30%	25%
partial char recall / precision	.58 / .56	.72 / .50
recall: invention	.36	.52
mischaracterization	.35	.47
OCR	.77	.84
miscounting	.62	.62
other	.65	.55

Table 6: Error decomposition before and after fusion (beds differ as labeled; both honest). The judge fixes exactly the deficits it was prompted to fix (§4).

Qualitative examples. Three EVAL archetypes, selected automatically (underline = gold union; **highlight** = v2 prediction; markings snapped to word boundaries for readability).

(a) *Judge suppresses a confident probe false alarm (gold-clean).* “Is the guy wearing specs?” → “Yes, the guy appears to be wearing spectacles. You can see a slight outline of frames around his eyes from this silhouette perspective.” The probe ensemble peaks at $p \geq .8$ on the hedged visual claim, but the D2 judge’s claim checklist passes it; the fused maximum stays below the 0.6 gate and v2 correctly abstains — worth a full 1.0 Cor.

(b) *Fusion catches an invention the probe missed.* “What is the color of the tool box?” → “Based on the image, **the toolbox is **yellow****. **It’s visible in the** upper left corner of the frame.” (no toolbox is identifiable in the image). The probe scores the span below threshold; the judge, answering the question itself first, fails to verify the object and marks it — the fused span matches the gold union (sample Cor 0.69). The trailing highlight shows the recall-push precision cost.

(c) *Failure mode: single-word mischaracterization.* “What is the color of the car outside the window...?” → “The car outside the window in front of the building door is white.” One annotator marked white ($\frac{1}{3}$); both legs pass it (a plausible color, a hard visual check), v2 predicts nothing, and the sample scores 0.0 — the archetype behind the still-low mischaracterization recall.

gold \ vote	<i>n</i>	inv.	mis.	OCR	mct.	oth.	missed
invention	1631	779	189	19	22	1	621
mischaracter.	1819	415	532	7	18	1	846
OCR	188	37	8	85	3	0	55
miscounting	331	19	44	2	121	0	145
other	152	37	15	0	2	2	96

Table 7: Gold-class $\times$ judge-majority-vote confusion, span level, EVAL ($n=1,200$ samples). Diagonal in bold.