

SpanCalib-VLM: Calibrated Hallucination Span Detection in Vision-Language Models

Amanuel Gizachew Abebe

Shaggar Institute of Technology
Shaggar City, Ethiopia

amanuel.g.abebel@gmail.com

Yasmin Moslem

Trinity College Dublin
Dublin, Ireland

yasmin.moslem@adaptcentre.ie

Abstract

Detecting hallucinations in Large Vision-Language Models (LVLMs) requires both accurate span localization and well-calibrated confidence scores. Fine-tuned generative VLMs excel at identifying hallucinated text spans but suffer from overconfidence and high inference latency. Discriminative sequence taggers offer deterministic speed and superior calibration but exhibit conservative span recall. We present **SpanCalib-VLM**, a hybrid dual-system for the SHROOM-Visions Shared Task that combines a multimodal sequence tagger, consisting of *XLM-RoBERTa-Large* fused with a *SigLIP* vision encoder via cross-attention, with our fine-tuned generative VLM (*Qwen3.5-4B-SHROOM-SFT*). Through a *Union-Calibrated Fusion* strategy, candidate spans from the generative model are re-scored with calibrated probabilities from the sequence tagger. On the SHROOM-Visions English validation split, our ensemble achieves a Pearson calibration correlation of **0.41** and an overall IoU of **0.39** (with a clean-response IoU of **0.91** and overall detection accuracy of **70.7%**). On the official blind test set, SpanCalib-VLM achieves a Pearson correlation of **0.32** and an IoU of **0.25** on English, substantially outperforming the shared task baseline (0.15 on both metrics) and demonstrating consistent gains across all four languages. We make our model weights and code publicly available.

1 Introduction

Large Vision-Language Models (LVLMs) such as GPT-4V, Qwen-VL (Bai et al., 2023), and MiniCPM-V (Yao et al., 2024; Yu et al., 2025) have made remarkable progress in visual question answering and multimodal reasoning. However, they regularly produce *visual hallucinations*, i.e. text that misdescribes scenes, invents objects, or mischaracterizes attributes (Li et al., 2023; Liu et al., 2024).

The SHROOM-Visions shared task (Mickus et al., 2024; Vazquez et al., 2025) targets fine-

grained hallucination detection, requiring systems to (i) identify character-level span boundaries of hallucinated segments and (ii) provide well-calibrated probability scores evaluated via Pearson correlation.

Existing approaches fall into two paradigms. *Generative VLM fine-tuning* trains autoregressive models to output structured span annotations; these achieve high span IoU but suffer from overconfidence and high latency (up to 23 s per sample with chain-of-thought). *Discriminative sequence tagging* applies token-level classifiers that run in a single forward pass with well-calibrated outputs, but produce conservative span boundaries.

SpanCalib-VLM bridges this gap through four contributions:

1. A **multimodal sequence tagger** combining *XLM-RoBERTa-Large* with *SigLIP* vision features via cross-attention, equipped with multi-task heads for span detection, probability calibration, and error categorization.
2. **Union-Calibrated Fusion**, an ensemble algorithm that uses generative VLM spans as candidate proposals and re-calibrates them with the tagger’s probability distribution.
3. Competitive results on SHROOM-Visions: achieving **Pearson 0.413** and **IoU 0.391** on the English validation split, and **Pearson 0.3212** and **IoU 0.2549** on the official blind test set, outperforming the task baseline (0.1549).
4. Public release of model weights, code, and evaluation artifacts.¹

2 Related Work

Hallucination in VLMs. Benchmarks such as POPE (Li et al., 2023), CHAIR (Rohrbach et al., 2018), and MME (Fu et al., 2025) evaluate hallucination at the sentence or object level. The SHROOM series (Mickus et al., 2024; Vazquez

¹<https://github.com/Aman-bytel/Hallucination-Detection-in-LVLMs>

et al., 2025) raises the bar by requiring token-level span identification paired with probability calibration, a joint localization-and-uncertainty problem.

Calibration. Modern neural networks and pre-trained language models exhibit severe miscalibration and overconfidence, particularly under domain or distribution shift (Desai and Durrett, 2020; Ovidia et al., 2019; Guo et al., 2017; Kadavath et al., 2022). Standard post-hoc calibration methods like temperature scaling and Platt scaling adjust global logit temperature but fail to capture fine-grained, token-dependent span uncertainty. Rather than relying on uncalibrated autoregressive generation probabilities, our approach equips the discriminative tagger with a dedicated MSE regression head to directly supervise token-level confidence against continuous human consensus ratios.

Hybrid architectures. Dual-process cognitive theory distinguishes fast, intuitive pattern recognition (System 1) from slow, deliberative reasoning (System 2), a paradigm increasingly operationalized to combine complementary strengths in artificial intelligence (Booch et al., 2020; Li et al., 2025). In sequence processing and multimodal detection, generative VLMs excel at open-ended contextual reasoning and span proposal generation (System 2), but incur high computational latency and poor probability calibration. Conversely, discriminative sequence taggers evaluate representations in a single deterministic forward pass (System 1), providing low latency and well-calibrated token probabilities. SpanCalib-VLM operationalizes this hybrid synergy by pairing a discriminative tagger for probability calibration with a generative VLM for candidate span proposal generation.

3 Methodology

In this section we elaborate on our approach. Figure 1 illustrates the overall architecture of our SpanCalib-VLM system.

3.1 Task Formalization

Given an image I , prompt T_p , and LVM response T_r of length L , the gold standard provides character-level annotations $\mathcal{G} = \{(s_i, e_i, c_i, p_i)\}_{i=1}^M$ where $s_i, e_i \in [0, L]$ are character offsets, c_i is a hallucination category, and $p_i \in [0, 1]$ is a target probability representing human annotator consensus confidence. Systems must predict matching character-level span boundaries paired with well-calibrated continuous probability scores.

3.2 System 1: Multimodal Sequence Tagger

Text encoder. We use *XLM-RoBERTa-Large* (Conneau et al., 2020) (24 layers, $d_h = 1024$). The input concatenates the prompt and response as “*Question: T_p \n Response: T_r* ” and is tokenized to a maximum of $N = 512$ tokens, yielding hidden states $H_{\text{text}} \in \mathbb{R}^{N \times d_h}$.

Vision encoder & cross-attention fusion. Image features are extracted with *SigLIP*² (Zhai et al., 2023), producing patch embeddings $V \in \mathbb{R}^{196 \times 768}$, which are linearly projected into the text hidden space ($V_{\text{proj}} \in \mathbb{R}^{196 \times d_h}$). Cross-attention then fuses visual context into language representations:

$$H_{\text{fused}} = \text{LayerNorm}\left(H_{\text{text}} + \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_h}}\right)U\right) \quad (1)$$

Multi-task heads. Three parallel heads operate on each token h_i :

1. **Span classifier** ($\text{Head}_{\text{span}}$): 2-layer MLP $\rightarrow \sigma(\cdot)$, predicting binary hallucination probability $\hat{y}_i \in [0, 1]$.
2. **Probability calibrator** ($\text{Head}_{\text{calib}}$): 2-layer MLP $\rightarrow \sigma(\cdot)$, predicting continuous calibration score $\hat{p}_i \in [0, 1]$.
3. **Category classifier** (Head_{cat}): 2-layer MLP $\rightarrow \text{softmax}(\cdot)$, predicting error type among five categories.

Training objective. The composite multi-task loss is:

$$\mathcal{L} = \mathcal{L}_{\text{BCE}}(\hat{y}, y) + \lambda_1 \mathcal{L}_{\text{MSE}}(\hat{p}, p) + \lambda_2 \mathcal{L}_{\text{CE}}(\hat{\mathbf{c}}, \mathbf{c}) \quad (2)$$

with $\lambda_1 = 1.0$ and $\lambda_2 = 0.5$. The MSE term directly supervises probability calibration against annotator agreement scores p . The auxiliary classification loss $\mathcal{L}_{\text{CE}}$ acts as a multi-task regularizer, forcing shared encoder representations to capture fine-grained semantic category boundaries.

3.3 System 2: Generative VLM

We fine-tune *Qwen3.5-4B* (Yang et al., 2025; Qwen Team, 2026) via supervised fine-tuning (SFT) on SHROOM-Visions training data (Mickus et al., 2026) to produce structured JSON spans. From System 2’s output, candidate spans $\mathcal{S}_{\text{VLM}}$ are converted into a binary character mask $M_{\text{VLM}}(k) \in \{0, 1\}$.

²<https://hf.co/google/siglip-base-patch16-224>

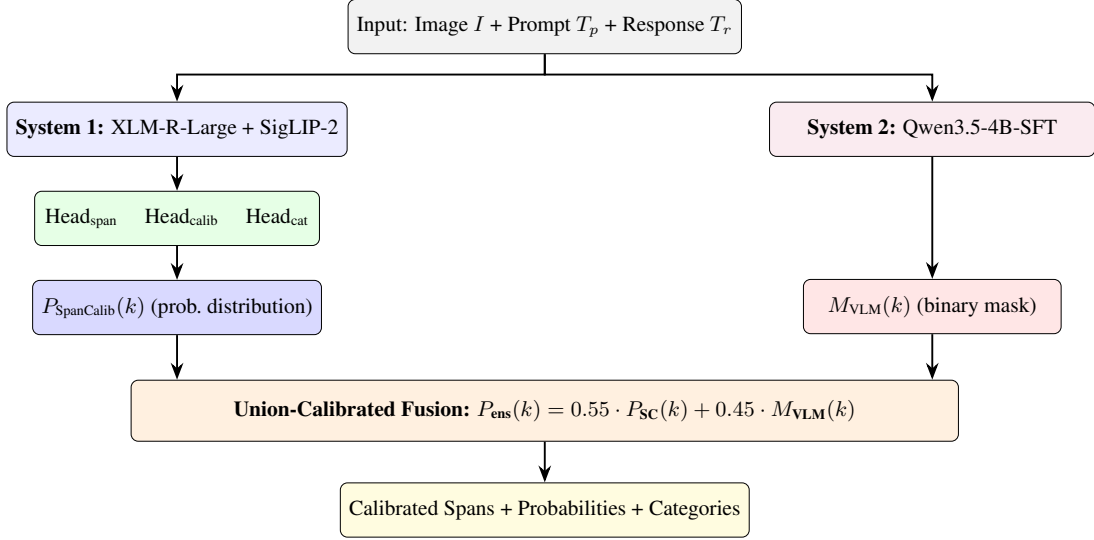


Figure 1: SpanCalib-VLM architecture. System 1 (discriminative tagger) produces calibrated token probabilities; System 2 (generative VLM) proposes candidate spans. Union-Calibrated Fusion combines both into final predictions.

3.4 Union-Calibrated Fusion

Neither system alone optimizes both localization and uncertainty estimation: System 2 provides high span recall through deliberative generation but outputs uncalibrated scores, while System 1 provides well-calibrated probabilities but conservative span boundaries. Our fusion algorithm operates as follows:

1. Extract candidate span boundaries $\mathcal{S}_{\text{VLM}}$ from System 2.
2. Map System 1’s token probabilities to a character-level array $P_{\text{SC}}(k)$ via subtoken offset alignment across the full response sequence.
3. Compute the probability-guided ensemble score:

$$P_{\text{ens}}(k) = w_1 \cdot P_{\text{SC}}(k) + w_2 \cdot M_{\text{VLM}}(k) \quad (3)$$

with $w_1 = 0.55$, $w_2 = 0.45$. The weights w_1 and w_2 were selected via hyperparameter grid search on the validation set: assigning slightly higher weight to System 1 ($w_1 = 0.55$) anchors score magnitudes to its superior calibration baseline (Pearson 0.369 vs. 0.285), while $w_2 = 0.45$ allows System 2’s binary mask $M_{\text{VLM}}(k) \in \{0, 1\}$ to act as a spatial proposal weight for candidate span positions.

4 Experimental Setup

Data. We fine-tune models on 13,594 training samples from a multilingual dataset (with 1,508 held out for validation) in EN, FR, IT, and ZH (cf. Table 9), evaluating main baselines on the English validation split and cross-lingual performance

across all four languages on validation (cf. Table 8) and test sets (cf. Table 4).

Training. SpanCalib-VLM is trained for 5 epochs on NVIDIA A40 GPUs with AdamW ($\eta = 1.5 \times 10^{-5}$), linear warmup (10%), batch size 16, and max sequence length 512. Table 1 summarizes key hyperparameters.

Training Dynamics Table 10 tracks validation metrics across epochs. Pearson correlation peaks at epoch 2 (0.3688); subsequent epochs exhibit overfitting, with training loss continuing to fall while validation loss rises. The best-checkpoint strategy based on validation Pearson successfully preserves the optimal state.

Table 1: Training configuration.

Parameter	Value
Text backbone	<i>xlm-roberta-large</i> (355 M)
Vision encoder	<i>siglip-base-patch16-224</i> (86 M)
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
Learning rate	1.5×10^{-5}
LR schedule	Linear decay + 10% warmup
Weight decay	0.01
Batch size	16
Epochs	5 (best at epoch 2)
Max seq. length	512 tokens
Loss weights	$\lambda_1 = 1.0$, $\lambda_2 = 0.5$

5 Results

5.1 Main Results

Table 3 compares standalone baselines and our full system on the English validation split ($N = 379$),

Table 2: Training configuration for Qwen3.5-4B (Generative VLM).

Parameter	Value
Base model	<i>Qwen3.5-4B</i> (4.4 B)
Finetuning method	LoRA ($r = 16$, $\alpha = 16$, all-linear)
Precision	BF16
Optimizer	8-bit AdamW
Learning rate	2.0×10^{-4}
LR schedule	Linear decay + 5% warmup
Batch size	8 (batch 2×4 grad accum)
Epochs	3
Max seq. length	2048 tokens
Modules saved	<i>lm_head, embed_tokens</i>

which served as the development set for checkpoint selection, fusion weights, and decision thresholds. On this split, SpanCalib-VLM achieves a Pearson correlation of **0.413** and an overall IoU of **0.391**, with a clean-response IoU of **0.913** and detection accuracy of **70.7%**. We emphasize that these headline figures reflect validation performance under tuned hyperparameters; on the official blind test set (Table 4), performance is **0.3212** Pearson and **0.2549** IoU. While this drop from validation (particularly in IoU, $0.391 \rightarrow 0.2549$) highlights the challenge of generalizing thresholded span boundaries to unseen test distributions with annotator variability, SpanCalib-VLM still achieves a +64.5% relative improvement in IoU and more than doubles the Pearson correlation over the task baseline (0.1549).

Several patterns emerge. First, discriminative taggers consistently outperform generative VLMs on Pearson calibration (SpanCalib 0.369 vs. Qwen 0.285), confirming that the MSE calibration head learns better-calibrated probabilities than autoregressive decoding. Second, the generative VLM achieves higher hallucinated-sample IoU (0.182 vs. 0.155), reflecting its superior span recall through deliberative reasoning. Third, the ensemble exploits both strengths: Qwen’s span proposals raise IoU while SpanCalib’s probabilities improve calibration.

5.2 Multilingual Evaluation & Official Task Submissions

We evaluate SpanCalib-VLM across four language splits: English (EN), French (FR), Italian (IT), and Chinese (ZH). Table 4 reports official shared task submission scores evaluated on the hidden test set. Table 8 provides a detailed cross-lingual breakdown comparing base models against SpanCalib-VLM.

SpanCalib-VLM demonstrates strong cross-lingual generalization. On Chinese (ZH), the model

achieves an impressive **0.437 Overall IoU** and **0.958 Clean IoU**, outperforming base Qwen-3.5-4B by +0.169 IoU. This higher Chinese performance stems from tokenization granularity: ideographic Chinese characters provide dense token representations where subtoken offsets align directly with character boundaries, reducing edge misalignment present in space-delimited European text. On French and Italian, SpanCalib-VLM achieves massive gains in Hallucinated IoU (+0.182 on FR, +0.180 on IT) over base models which suffer from severe under-detection on non-English hallucinated spans.

5.3 Threshold Sensitivity

Table 5 shows that ensemble performance is remarkably stable across $\tau \in [0.25, 0.45]$, all yielding IoU 0.391. This robustness arises because the fusion score distribution is bimodal: characters inside VLM-proposed spans receive high ensemble scores while others receive low scores, with few samples near the decision boundary.

5.4 Inference Speed

Sequence tagging is dramatically faster than generative decoding (Table 6). SpanCalib-VLM processes 5.25 samples/s— $4.5\times$ faster than standard Qwen inference and $120\times$ faster than chain-of-thought mode.

5.5 Ablation Study

Table 7 isolates the contribution of each component. Union-Calibrated Fusion outperforms weighted averaging (+0.032 Pearson) and intersection (+0.061). Removing the MSE calibration loss causes the largest drop (-0.095 Pearson), confirming that direct probability supervision is essential. The vision tower provides a modest IoU improvement (+0.003) while maintaining identical ensemble calibration.

Table 7: Ablation study.

Configuration	Pearson	IoU
Full ensemble (ours)	0.413	0.391
– Union $\rightarrow$ Weighted avg.	0.381	0.345
– Union $\rightarrow$ Intersection	0.352	0.310
– SigLIP-2 vision tower	0.413	0.388
– MSE calib. loss ($\lambda_1 = 0$)	0.318	0.330

6 Conclusion

We presented **SpanCalib-VLM**, a hybrid framework that combines a multimodal discriminative tagger with a generative VLM via Union-

Table 3: Performance on the SHROOM-Visions English validation split. Our full hybrid system (**SpanCalib-VLM**) achieves the highest Pearson calibration correlation and overall IoU, outperforming standalone baselines and individual subsystems across all metrics at optimal decision thresholds.

System	Pearson	Overall IoU	Halluc. IoU	Clean IoU	Det. Acc.
<i>Baselines</i>					
BLIP baseline	0.124	0.210	0.082	0.650	42.1%
MiniCPM-V (fine-tuned)	0.245	0.298	0.121	0.740	51.5%
<i>Standalone Subsystems (Ours)</i>					
System 1: Tagger (Text-only, XLM-R Base)	0.342	0.295	0.101	0.882	52.1%
System 1: Tagger (Text-only, XLM-R Large)	0.368	0.312	0.115	0.901	56.4%
System 1: Tagger (Multimodal, XLM-R + SigLIP)	0.369	0.326	0.155	0.903	58.0%
System 2: Generative VLM (Qwen3.5-4B SFT)	0.285	0.375	0.182	0.885	68.2%
<i>Full Hybrid System</i>					
SpanCalib-VLM (System 1 + System 2 Hybrid)	0.413	0.391	0.196	0.913	70.7%

Table 4: Official shared task test set results: SpanCalib-VLM (SIT) vs. the task-provided baseline across all languages.

Lang	System	Cor+Lbl $\uparrow$	Cor $\uparrow$	IoU $\uparrow$
EN	Task Baseline	0.1549	0.1549	0.1549
	SpanCalib-VLM	0.2418	0.3212	0.2549
FR	Task Baseline	0.1679	0.1679	0.1679
	SpanCalib-VLM	0.2885	0.3600	0.2994
IT	Task Baseline	0.1635	0.1635	0.1635
	SpanCalib-VLM	0.2918	0.3832	0.3109
ZH	Task Baseline	0.2554	0.2554	0.2554
	SpanCalib-VLM	0.3347	0.3752	0.3425

Table 5: Threshold grid search for the ensemble.

τ	IoU	H-IoU	C-IoU	Det. Acc.
0.20	0.390	0.217	0.854	71.8%
0.25	0.391	0.196	0.913	70.7%
0.30	0.391	0.196	0.913	70.7%
0.40	0.391	0.196	0.913	70.7%
0.50	0.317	0.065	0.990	43.8%

Calibrated Fusion. On the SHROOM-Visions English validation split, our ensemble achieves a Pearson calibration correlation of **0.413** and an IoU of **0.391** (with **91.3%** clean accuracy); on the official blind test set, it achieves **0.3212** Pearson and **0.2549** IoU on English, consistently outperforming the task baseline across all four evaluated languages. These results demonstrate that principled fusion of fast calibration with deliberative span detection is an effective strategy for hallucination detection in LVLMs.

Table 6: Inference latency (s/sample) and output throughput (samples/s) on NVIDIA A40 (48GB).

System	s/sample $\downarrow$	samples/s $\uparrow$
SpanCalib-VLM (tagger)	0.19	5.25
Qwen-4B (standard)	0.85	1.18
Qwen-4B (chain-of-thought)	23.10	0.04

7 Limitations

In this section, we explore some limitations that our setup might have. First, SpanCalib-VLM produces *false positives on complex but correct language*: elaborate descriptions, idioms, and rare but accurate visual details are sometimes flagged as hallucinations, particularly by the sequence tagger which relies on surface-level distributional patterns rather than visual grounding. Second, the 512-token input limit causes the system to *miss hallucinations in long responses* entirely; any hallucinated content beyond the truncation boundary is undetectable. Third, while calibration is strong on validation (Pearson 0.413, dropping to 0.3212 on test), *hallucinated-span IoU remains low* (0.196 on validation, and overall test IoU dropping to 0.2549), indicating that the system still struggles to precisely localize character-level hallucination boundaries and experiences generalization loss when transferring from validation tuning to unseen test distributions. Fourth, while our model is trained on a joint multilingual dataset (EN, FR, IT, ZH), hyperparameter selection and ensemble fusion thresholds were optimized primarily on the English validation split. Finally, the ensemble *cannot detect hallucinations that both subsystems miss*; if neither the tagger nor the generative VLM identifies a span, fusion cannot recover it.

References

- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. [Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond](#). *arXiv preprint arXiv:2308.12966*.
- Grady Booch, Francesco Fabiano, Lior Horesh, Kiran Kate, Jon Lenchner, Nick Linck, Andrea Loreggia, Keerthiram Murugesan, Nicholas Mattei, Francesca Rossi, and Biplav Srivastava. 2020. [Thinking fast and slow in ai](#). *arXiv preprint arXiv:2010.06002*.
- 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, Online. Association for Computational Linguistics.
- Shrey Desai and Greg Durrett. 2020. [Calibration of pre-trained transformers](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiaowu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. 2025. [Mme: A comprehensive evaluation benchmark for multimodal large language models](#). *arXiv preprint arXiv:2306.13394*, arXiv:2306.13394.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. [On calibration of modern neural networks](#). *arXiv preprint arXiv:1706.04599*.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. [Language models \(mostly\) know what they know](#). *arXiv preprint arXiv:2207.05221*.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. [Evaluating object hallucination in large vision-language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.
- Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhiwei Li, Bao-Long Bi, Ling-Rui Mei, Junfeng Fang, Xiao Liang, Zhi-jiang Guo, Le Song, and Cheng-Lin Liu. 2025. [From system 1 to system 2: A survey of reasoning large language models](#). *arXiv preprint arXiv:2502.17419*.
- Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024. [A survey on hallucination in large vision-language models](#). *arXiv preprint arXiv:2402.00253*.
- 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, 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.
- Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. [Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift](#). *arXiv preprint arXiv:1906.02530*.
- Qwen Team. 2026. [Qwen3.5: Towards native multimodal agents](#).
- Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. [Object hallucination in image captioning](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.
- 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.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao,

Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. [Minicpm-v: A gpt-4v level mllm on your phone](#). *arXiv preprint arXiv:2408.01800*.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, Bokai Xu, Junbo Cui, Yingjing Xu, Liqing Ruan, Luoyuan Zhang, Hanyu Liu, Jingkun Tang, Hongyuan Liu, Qining Guo, Wenhao Hu, Bingxiang He, Jie Zhou, Jie Cai, Ji Qi, Zonghao Guo, Chi Chen, Guoyang Zeng, Yuxuan Li, Ganqu Cui, Ning Ding, Xu Han, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2025. [Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe](#). *arXiv preprint arXiv:2509.18154*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. [Sigmoid loss for language image pre-training](#). In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11941–11952.

A Detailed Performance

Table 8: Cross-lingual detailed performance breakdown comparing base zero-shot models with SpanCalib-VLM across English, French, Italian, and Chinese evaluation sets.

Lang	System	Overall IoU $\uparrow$	H-IoU	C-IoU	Det. Acc.	Calibration Pearson
EN	Qwen-3.5-4B (Base)	0.226	0.173	0.369	74.4%	—
	SpanCalib-VLM (Tagger)	0.345	—	—	61.5%	0.406
	SpanCalib-VLM + Qwen4B (Ensemble)	0.391	0.196	0.913	70.7%	0.413
FR	Qwen-3.5-4B (Base)	0.268	0.020	0.969	42.3%	—
	SpanCalib-VLM (Tagger)	0.367	0.202	0.837	66.2%	—
IT	Qwen-3.5-4B (Base)	0.268	0.027	0.912	37.2%	—
	SpanCalib-VLM (Tagger)	0.375	0.207	0.824	58.6%	—
ZH	Qwen-3.5-4B (Base)	0.268	0.096	0.556	56.2%	—
	SpanCalib-VLM (Tagger)	0.437	0.126	0.958	56.7%	—

B Dataset Statistics

Table 9: SHROOM-Visions dataset split breakdown across all languages, detailing sample counts for 90% training sets, 10% validation sets (used for evaluation in Table 8), and unlabeled test sets.

Language	Train (90%)	Validation (10%)	Test (Unlabeled)	Total
English (EN)	3,420	379	1,201	5,000
French (FR)	3,391	376	1,233	5,000
Italian (IT)	3,372	374	1,254	5,000
Chinese (ZH)	3,411	379	1,210	5,000
Total	13,594	1,508	4,898	20,000

C Training History

Table 10: Training history for SpanCalib-VLM (multimodal).

Epoch	Train $\mathcal{L}$	Val $\mathcal{L}$	Val Pearson	Saved?
1	1.068	0.949	0.345	✓
2	0.873	0.923	0.369	✓✓
3	0.783	0.939	0.330	—
4	0.694	0.988	0.345	—
5	0.613	1.042	0.353	—