Knowing When to Defer: Gradient-Free Abstention for Vision-Language Models in Surgical VQA via Neighbourhood Disagreement

No Author Given

Abstract. Off-the-shelf vision-language models (VLMs) are increasingly used as frozen models for surgical visual question answering (VQA), where task-specific fine-tuning is often impractical due to limited labelled data and compute. However, they frequently answer surgical questions with high confidence despite being incorrect, making them unsafe in high-stakes settings where a model should defer under uncertainty. To this end, we propose Neighbourhood Disagreement, a novel gradient-free abstention method for frozen VLMs in surgical VQA. Our method samples local experts by applying random Gaussian perturbations to the frozen model weights and uses disagreement among the resulting perturbed experts as an epistemic uncertainty signal. The perturbation radius σ acts as a temperature-like control in weight space. Small radii preserve overconfident agreement, intermediate radii expose useful disagreement while retaining task competence, and overly large radii trigger an accuracy cliff as expert behaviour collapses. The neighbourhood disagreement improves selective prediction by identifying cases where the frozen model should defer. Across three surgical VQA datasets, CoPESD, Cholec80-VQA, and SSG-VQA, we find a narrow transferable operating regime for Qwen2.5-VL-7B around $\sigma^* \approx 0.004$, near the edge of competence. This regime consistently improves calibration by reducing Expected Calibration Error from 0.27/0.28/0.30 to 0.04/0.07/0.04 while also improving selective prediction, cutting the Area under risk coverage from 0.56/0.41/0.53 to 0.35/0.37/0.37 on CoPESD/Cholec80-VQA/SSG-VQA. Our results on 4 VLMs show that weight-space neighbourhood disagreement provides a strong uncertainty signal for frozen VLMs in surgical VQA.

1 Introduction and Background

Vision–language models (VLMs) [12, 3, 18] are increasingly being explored for surgical visual question answering (VQA), where models are expected to interpret operative images or videos and answer clinically meaningful questions about anatomy, instruments, procedural steps, and surgical actions [14, 16, 2, 5]. This is desirable because a single frozen foundation model can be deployed across tasks without task-specific training, reducing the need for costly annotation and repeated fine-tuning. This is particularly important in surgery, where labelled data are scarce, source datasets may be inaccessible because of privacy constraints, and model updates may be impractical in deployment settings [13, 11, 15, 6].

However, accuracy alone is insufficient for high-stakes surgical VQA. A model that answers incorrectly but with high confidence is more dangerous than an

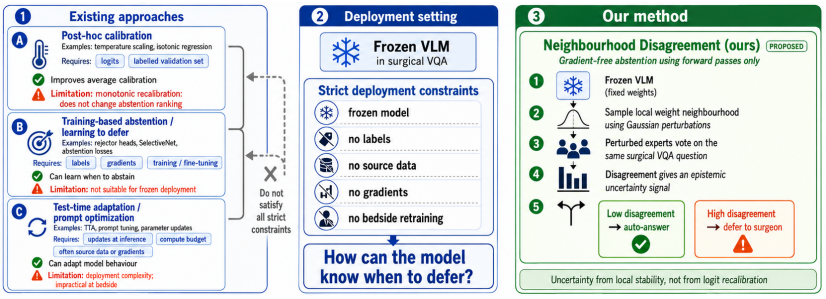


Fig. 1: Positioning relative to existing calibration and abstention approaches

uncertain model. Recent VLMs often produce confident predictions even when visual evidence is ambiguous, domain-specific, or outside the pretraining distribution. In surgical VQA, this problem is especially pronounced because visually similar instruments, occluded anatomy, subtle procedural cues, and strong language priors can cause a frozen VLM to select plausible but incorrect answers [10, 1, 17, 4]. A reliable surgical VQA system should not only answer questions, but also identify when its answer should not be trusted and should instead be deferred to a surgeon. This shifts the focus from calibration alone to abstention under deployment constraints. Standard post-hoc calibration methods, such as temperature scaling, can improve average confidence alignment when logits and labelled validation data are available, but they are limited for selective prediction as monotonic recalibration changes confidence values without changing the ranking of examples by reliability [8, 9]. Thus, it cannot determine which cases a frozen model should answer or defer. Other approaches require fine-tuning, prompt optimization, test-time updates, source data, or additional annotation, which are often impractical for surgical VQA. We therefore ask whether a useful abstention signal can be obtained for frozen VLMs using only forward passes.

This paper proposes *Neighbourhood Disagreement*, a gradient-free abstention method that probes the local weight neighbourhood of a frozen VLM. We sample perturbed experts by applying random Gaussian perturbations to the model weights and use their disagreement as an epistemic uncertainty signal. Concentrated votes indicate stable predictions suitable for automatic answering, whereas dispersed votes indicate local instability and support deferral. The perturbation radius σ acts as a temperature-like control in weight space. Small radii preserve overconfident agreement, intermediate radii expose useful disagreement while retaining task competence, and overly large radii cause an accuracy cliff as expert behaviour collapses. Useful abstention therefore emerges near the model's *edge of competence*, where the ensemble remains accurate but no longer overconfident. Unlike conventional logit-space calibration, we derive uncertainty from prediction stability under local weight perturbations. Across 3 surgical datasets, increasing σ reduces overconfidence until a narrow operating regime, after which accuracy rapidly collapses. This regime reduces expected calibration error while improving accuracy. More importantly, Neighbourhood Disagreement improves selective prediction by re-ranking cases according to reliability, enabling uncertain surgical questions to be deferred rather than answered automatically.

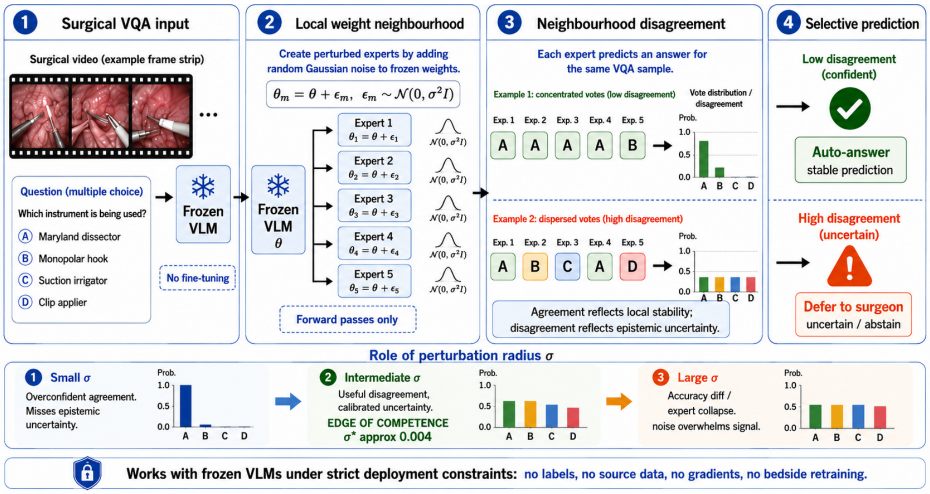


Fig. 2: Overview of our proposed Neighbourhood Disagreement

Our primary contributions are threefold: First, we introduce Neighbourhood Disagreement, a novel gradient-free abstention method for frozen VLMs in surgical VQA based on local weight-perturbed expert sampling. Second, we show that σ provides a temperature-like control over expert agreement, revealing an operating regime near the *edge of competence*. Third, we demonstrate that Neighbourhood Disagreement improves uncertainty calibration and selective prediction without fine-tuning or labelled validation sets.

2 Methodology

Problem Setup: We consider multiple-choice surgical visual question answering (VQA), where each sample consists of a surgical visual input x_i, a natural-language question q_i, a set of K candidate answers $\mathcal{A}_i = a_i^1, \ldots, a_i^K$, and a ground-truth answer $y_i \in 1, \ldots, K$. A frozen vision-language model (VLM) f_{θ_0}, with parameters θ_0, is prompted to select one answer option. The goal is not only to predict the correct answer, but also to produce a confidence estimate whose value reflects the probability of correctness. Standard calibration methods usually require access to logits or model gradients. In contrast, our setting assumes a frozen VLM and prohibits fine-tuning, gradient updates, and logit modification at inference time. We therefore estimate uncertainty from the behaviour of local ensemble of perturbed models around θ_0. The central hypothesis is that the local weight neighbourhood of a frozen VLM contains useful functional diversity, and that disagreement among nearby experts provides a practical uncertainty signal.

Weight-Space Temperature: We adopt a Neural Thicket [7] as an ensemble of M local experts sampled around the frozen VLM. For a perturbation radius σ, each expert $m \in 1, \ldots, M$ is obtained by applying random Gaussian perturbations to the frozen model weights: $\theta_m = \theta_0 + \Delta_m, \Delta_m \sim \mathcal{N}(0, \sigma^2 I)$. Here, σ controls the scale of the perturbation and therefore the spread of the local expert ensemble. We interpret σ as a *weight-space temperature*. Small values of σ produce experts close to the original model, resulting in high member agreement

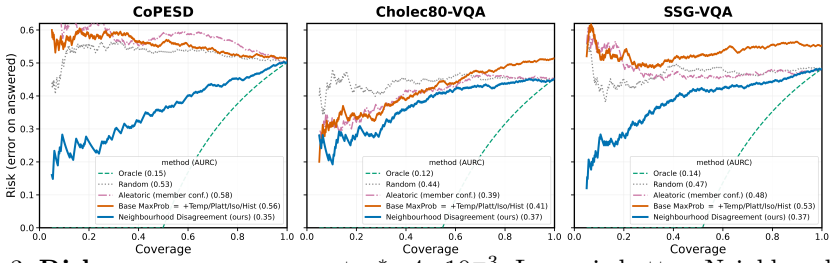


Fig. 3: **Risk–coverage curves** at σ^*=4×10^{-3}. Lower is better. Neighbourhood Disagreement (ours) lies below every deployable baseline at all coverages.

and potentially overconfident predictions. Larger values of σ induce greater diversity among experts, exposing disagreement and reducing confidence. However, if σ becomes too large, the perturbations destroy useful model behaviour and accuracy collapses. Calibration is therefore expected to occur in an intermediate window near the model's *edge of competence*.

Epistemic Uncertainty Estimation: For each surgical VQA sample i, every perturbed expert f_{θ_m} receives the same visual input, question, and answer options. The expert outputs a predicted answer: $\hat{y}_i^{(m)} \in 1, \ldots, K$. The M predictions are then converted into vote counts: $n_i^k = \sum_{m=1}^{M} \mathbf{1}!\left[\hat{y}_i^{(m)} = k\right], \quad k \in 1, \ldots, K$. We do not require VLM token probabilities or calibrated softmax scores. Instead, we construct an empirical predictive distribution from the ensemble votes. To avoid degenerate zero-probability estimates, we apply Dirichlet smoothing: $q_i^k = \frac{\alpha + n_i^k}{K\alpha + M}$ where α is the smoothing parameter. The neighbourhood prediction is the option with maximum smoothed vote mass: $\hat{y}_i = \arg\max_k q_i^k$. The corresponding confidence is defined as: $\mathrm{c}_i = \max_k q_i^k$. This confidence measures the concentration of the local expert vote distribution. If most experts agree, c_i is high. If votes are dispersed across candidate answers, c_i is low. Thus, uncertainty can be estimated as the complement of confidence: $U_i = 1 - \max_k q_i^k$ which corresponds to a variation-ratio uncertainty score.

Perturbation Radius Selection: We evaluate a grid of perturbation radii: $\Sigma = \sigma_1, \sigma_2, \ldots, \sigma_L$. For each $\sigma \in \Sigma$, we construct a neighbourhood of perturbed experts, obtain predictions and uncertainty scores. The radius σ controls the amount of local functional diversity. Small values preserve agreement with the frozen model, intermediate values expose useful epistemic disagreement, and overly large values destroy task-relevant behaviour. A radius σ^* is considered well calibrated only if it simultaneously reduces the signed gap and achieves low expected calibration error (ECE). Once σ^* has been selected, the model produces predictions and uncertainty estimates using only perturbed-expert votes.

Edge of Competence: A key property of Neighbourhood Disagreement is that useful uncertainty is bounded by an accuracy cliff as shown in Fig. 4. Increasing σ initially reduces overconfidence by inducing disagreement among nearby experts. However, beyond a critical radius, perturbations no longer produce meaningful local experts. Their predictions become unstable or random, and accuracy collapses toward chance. We refer to the useful region immediately before this

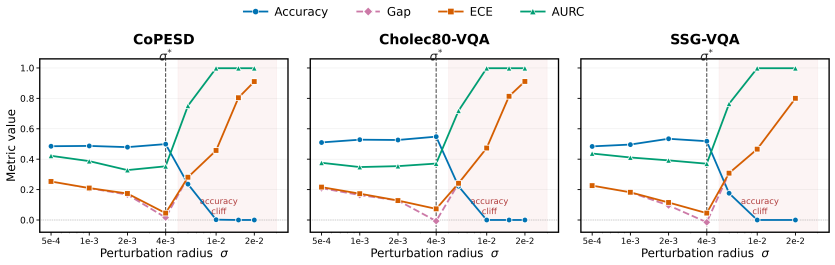


Fig. 4: **Weight-space temperature.** Effect of varying perturbation radius σ for frozen Qwen2.5-VL-7B. Calibration (Gap, ECE) and selective prediction (AURC) improve to a minimum at $\sigma^* \approx 4 \times 10^{-3}$ with accuracy preserved, then collapse past the accuracy cliff ($\sigma \geq 6 \times 10^{-3}$, shaded).

collapse as the *edge of competence*. A useful operating radius should satisfy three conditions. It should: (a) reduce overconfidence, (b) maintain accuracy close to the unperturbed or low-perturbation ensemble, and (c) avoid the high-σ regime where accuracy collapses. This distinguishes meaningful epistemic disagreement from uncertainty caused by random degradation.

3 Experiments and Results

Datasets, Settings, and Baselines: We evaluate on three complementary surgical VQA datasets: **(i) CoPESD**, which tests fine-grained endoscopic submucosal dissection motion understanding; **(ii) Cholec80-VQA**, which tests laparoscopic cholecystectomy phase and tool-related questions; and **(iii) SSG-VQA**, which tests scene-graph-grounded spatial and action reasoning over instruments and anatomy. We report Expected Calibration Error (ECE), confidence–accuracy gap, Area Under Risk–Coverage curve (AURC) and compare with calibration baselines (Temperature, Platt, Isotonic, Histogram binning). To assess whether neighbourhood disagreement is useful for deployment, we evaluate selective prediction. Given an uncertainty score U_i, the model answers only samples whose uncertainty is below a threshold τ, abstaining on the rest: $\mathcal{S}_\tau = \{i : U_i \leq \tau\}$. Coverage is defined as: $\text{Coverage}_\tau = \frac{|\mathcal{S}_\tau|}{N}$, and selective accuracy is: $\text{Acc}_\tau = \frac{1}{|\mathcal{S}_\tau|} \sum_{i \in \mathcal{S}_\tau} \text{acc}_i$. Equivalently, selective risk is: $\text{Risk}_\tau = 1 - \text{Acc}_\tau$.

Neighbourhood Disagreement Improves Abstention Fig. 3 shows that perturbed-expert disagreement improves the risk–coverage trade-off over the model's native confidence. Fig. 4 further shows that at $\sigma = 0.004$, AURC decreases from 0.42 to 0.35 on CoPESD and from 0.44 to 0.37 on SSG-VQA, while ECE drops from 0.25 to 0.04, 0.22 to 0.07, and 0.23 to 0.04 on CoPESD, Cholec80-VQA, and SSG-VQA, respectively. Accuracy is preserved or improved at the same operating radius. Per-category results in Tables 1–2 further show that the effect generalizes across surgical phase, instrument, spatial, and action-oriented questions, indicating that local weight-neighbourhood disagreement provides a meaningful abstention signal across surgical reasoning settings.

Post-hoc Calibration Cannot Substitute for Abstention We compare Neighbourhood Disagreement with confidence and post-hoc calibration baselines, including MaxP, temperature scaling, Platt scaling, isotonic regression,

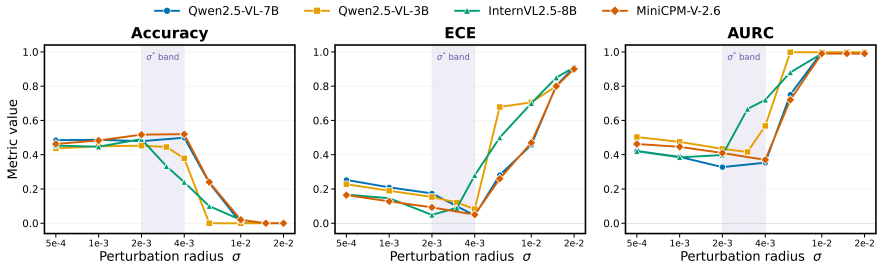


Fig. 5: **Architecture generalisation.** Accuracy, ECE, and AURC vs. perturbation radius σ for 4 VLMs on CoPESD. All architectures exhibit the same calibrate-then-cliff behaviour, but the calibrating σ^* and the accuracy cliff vary.

Table 1: Per-category results on CoPESD & Cholec80-VQA. Acc.↑,ECE/AURC↓

Dataset	Group / Category	Metric	0.0005	0.001	0.002	0.004	0.006	0.01	0.015	0.02
CoPESD	Forceps	Acc.	0.464	0.468	0.432	0.500	0.189	0.000	0.000	0.000
		Conf.	0.770	0.702	0.647	0.515	0.501	0.447	0.806	0.911
		Gap	+0.306	+0.234	+0.215	+0.015	+0.311	+0.447	+0.806	+0.911
		ECE	0.316	0.310	0.317	0.258	0.316	0.447	0.806	0.911
		AURC	0.437	0.459	0.446	0.470	0.798	0.953	0.953	0.953
	Knife	Acc.	0.462	0.464	0.452	0.454	0.260	0.003	0.000	0.000
		Conf.	0.704	0.664	0.610	0.486	0.521	0.466	0.805	0.911
		Gap	+0.242	+0.201	+0.158	+0.032	+0.261	+0.463	+0.805	+0.911
		ECE	0.261	0.255	0.222	0.227	0.268	0.463	0.805	0.911
		AURC	0.488	0.465	0.449	0.475	0.701	0.986	0.986	0.986
Cholec80-VQA	Surgical Phase	Acc.	0.356	0.380	0.380	0.468	0.240	0.000	0.000	0.000
		Conf.	0.587	0.553	0.524	0.464	0.485	0.467	0.812	0.912
		Gap	+0.231	+0.173	+0.144	-0.004	+0.245	+0.467	+0.812	+0.912
		ECE	0.243	0.191	0.144	0.119	0.247	0.467	0.812	0.912
		AURC	0.633	0.623	0.567	0.543	0.680	0.996	0.996	0.996
	Surgical Instruments	Acc.	0.664	0.676	0.672	0.628	0.204	0.000	0.000	0.000
		Conf.	0.852	0.831	0.784	0.618	0.439	0.478	0.813	0.910
		Gap	+0.188	+0.155	+0.112	-0.010	+0.235	+0.478	+0.813	+0.910
		ECE	0.218	0.180	0.136	0.039	0.241	0.479	0.813	0.910
		AURC	0.274	0.253	0.281	0.301	0.770	0.996	0.996	0.996

and histogram binning. As shown in Table 3, these baselines quickly lose coverage as the selective-accuracy target increases: on CoPESD and SSG, even the strongest calibrated baselines provide at most 1% coverage at the 70% target. In contrast, Neighbourhood Disagreement achieves the highest coverage across all datasets, covering 33% of CoPESD, 30% of Cholec80-VQA, and 22% of SSG-VQA at the 70% target, while retaining non-zero coverage even at 80%. These results show that post-hoc calibration can adjust confidence values, but does not necessarily produce the reliability ordering needed for safe abstention.

Perturbation radius controls uncertainty at the edge of competence: Increasing σ initially reduces overconfidence by exposing disagreement among nearby perturbed experts (see Fig. 4). A narrow operating regime around $\sigma^* \approx 0.004$ yields the best calibration behaviour while preserving accuracy. Beyond this point, performance collapses sharply, with an accuracy cliff appearing at $\sigma \geq 0.006$. This shows that σ acts as a temperature-like control in weight space, but useful uncertainty only emerges before perturbed experts lose task competence.

Safe Deployment Through Target-Accuracy Coverage To assess deployment utility, we measure the fraction of each dataset that can be answered while meeting a target selective accuracy. Table 3 shows that Neighbourhood Disagreement achieves substantially higher safe coverage than confidence and calibration

Table 2: **SSG-VQA per-category results** for Qwen2.5-VL-7B

Category	Metric	0.0005	0.001	0.002	0.004	0.006	0.01
Action on anatomy	Acc.	0.417	0.375	0.417	0.375	0.167	0.000
	Conf.	0.673	0.609	0.554	0.450	0.479	0.448
	Gap	+0.257	+0.234	+0.137	+0.075	+0.312	+0.448
	ECE	0.257	0.234	0.206	0.187	0.312	0.448
	AURC	0.596	0.559	0.615	0.725	0.678	0.979
Anatomy at location	Acc.	0.409	0.433	0.470	0.445	0.146	0.000
	Conf.	0.672	0.647	0.608	0.479	0.457	0.471
	Gap	+0.264	+0.214	+0.139	+0.034	+0.311	+0.471
	ECE	0.264	0.214	0.188	0.076	0.311	0.471
	AURC	0.463	0.495	0.460	0.411	0.793	0.994
Object at location	Acc.	0.567	0.567	0.592	0.600	0.192	0.000
	Conf.	0.743	0.733	0.680	0.539	0.499	0.463
	Gap	+0.176	+0.166	+0.088	-0.061	+0.307	+0.463
	ECE	0.202	0.190	0.142	0.081	0.307	0.463
	AURC	0.446	0.400	0.356	0.306	0.762	0.992
Tool action	Acc.	0.558	0.512	0.605	0.581	0.256	0.000
	Conf.	0.729	0.686	0.654	0.541	0.480	0.465
	Gap	+0.171	+0.174	+0.049	-0.040	+0.224	+0.465
	ECE	0.252	0.220	0.110	0.113	0.259	0.465
	AURC	0.414	0.319	0.277	0.208	0.572	0.977
Tool at location	Acc.	0.538	0.590	0.615	0.538	0.128	0.000
	Conf.	0.813	0.766	0.677	0.537	0.510	0.465
	Gap	+0.274	+0.176	+0.062	-0.001	+0.382	+0.465
	ECE	0.318	0.177	0.153	0.153	0.382	0.465
	AURC	0.334	0.343	0.324	0.394	0.903	0.974
Tool for anatomy	Acc.	0.487	0.513	0.487	0.538	0.256	0.000
	Conf.	0.750	0.709	0.619	0.537	0.521	0.482
	Gap	+0.262	+0.196	+0.132	-0.001	+0.265	+0.482
	ECE	0.302	0.206	0.270	0.077	0.265	0.482
	AURC	0.395	0.354	0.382	0.281	0.695	0.974

baselines. At the 70% target, it covers 33%, 30%, and 22% of CoPESD, Cholec80-VQA, and SSG-VQA, compared with at most 1%, 10%, and 1% for the strongest baselines, thereby providing a controllable safety-coverage trade-off by answering locally stable cases and deferring uncertain ones.

Cross-procedure Transfer of the Operating Radius We examine whether the perturbation regime transfers across surgical VQA settings. Fig. 4 shows a consistent trend: performance improves up to an intermediate regime around $\sigma^* \approx 0.004$, then sharply collapses near $\sigma \geq 0.006$. This suggests that, within a fixed VLM family, σ can be selected from a narrow operating range and reused across related surgical datasets. However, Fig. 5 shows that the optimal radius varies across architectures, indicating that σ^* is model-family dependent and should be lightly re-selected when changing VLM backbones.

Neighbourhood Disagreement Requires a Competence Floor Fig. 4 highlights a key failure mode that it is useful only while perturbed experts remain task-competent. Small perturbations preserve the base model's overconfident behaviour, while the intermediate regime around $\sigma = 0.004$ improves calibration and selective prediction without sacrificing accuracy. However, when σ becomes too large, performance collapses: at $\sigma = 0.006$ accuracy drops sharply, and for $\sigma \geq 0.01$ accuracy approaches zero while confidence remains non-negligible. Thus, the method does not create task knowledge; it is informative only when perturbations expose instability around a still-competent predictor.

Ablation We ablate the number of perturbed experts M used to estimate neighbourhood disagreement. Table 4 shows that increasing M improves both stability and abstention quality across datasets. Beyond $M \approx 20$, improvements slowly saturate, suggesting that larger ensembles mainly refine the estimate rather than

Table 3: Selective-prediction coverage (%) at target accuracies (%)

	CoPESD				Cholec80				SSG			
Method	≥50	≥60	≥70	≥80	≥50	≥60	≥70	≥80	≥50	≥60	≥70	≥80
MaxP	0	0	0	0	87	43	9	5	36	1	0	0
Platt	85	7	1	0	87	43	9	5	36	1	0	0
Iso	94	0	0	0	85	42	7	5	54	2	1	1
Temp	0	0	0	0	87	43	9	5	36	1	0	0
Hist	88	6	0	0	28	16	10	0	15	2	0	0
Ours	**99**	**61**	**33**	**7**	**100**	**51**	**30**	**10**	**100**	**45**	**22**	**8**

Table 4: Ablation over ensemble size M. Acc. ↑; ECE ↓; AURC ↓.

	CoPESD			Cholec80			SSG		
M	Acc.	ECE	AURC	Acc.	ECE	AURC	Acc.	ECE	AURC
1	.39±.06	.11±.06	.69±.06	.43±.07	.08±.06	.63±.07	.41±.06	.10±.05	.59±.07
2	.39±.06	.09±.04	.60±.06	.43±.06	.08±.05	.55±.06	.41±.05	.08±.05	.52±.06
3	.42±.04	.09±.04	.53±.06	.45±.05	.07±.04	.52±.06	.44±.05	.06±.04	.48±.06
5	.45±.04	.07±.03	.48±.06	.47±.03	.06±.02	.47±.04	.46±.05	.06±.03	.45±.05
8	.46±.02	.06±.02	.44±.04	.50±.03	.05±.02	.43±.04	.49±.03	.05±.02	.41±.03
10	.47±.02	.06±.01	.41±.03	.51±.02	.04±.01	.42±.03	.49±.02	.05±.01	.41±.03
15	.48±.01	.06±.01	.39±.03	.52±.02	.05±.02	.39±.02	.51±.02	.04±.01	.38±.02
20	.48±.01	.06±.02	.38±.02	.53±.02	.05±.02	.38±.01	.51±.02	.04±.01	.39±.02
25	.48±.01	.06±.01	.36±.01	.54±.01	.05±.01	.38±.01	.52±.01	.05±.01	.37±.01
30	.50±.01	.05±.01	.35±.01	.55±.01	.07±.01	.37±.02	.52±.01	.05±.01	.37±.01

change the signal. We use $M = 30$ which provides the most stable estimates and achieves the best or near-best AURC across datasets.

4 Conclusion

We presented Neighbourhood Disagreement, a gradient-free abstention method for frozen VLMs in surgical VQA that probes the local weight neighbourhood using random Gaussian perturbations and estimates uncertainty from disagreement among perturbed experts, requiring only forward passes and no fine-tuning, source data, labelled validation sets, logit access, or inference-time gradients. Across CoPESD, Cholec80-VQA, and SSG-VQA datasets, we show that the perturbation radius σ acts as a temperature-like control in weight space. Small perturbations preserve overconfident behaviour while large perturbations cause an accuracy cliff. A narrow regime around $\sigma^* \approx 0.004$ for Qwen2.5-VL-7B consistently improves calibration and selective prediction, enabling the model to defer uncertain cases. Our results highlight that abstention requires more than post-hoc confidence calibration. While calibration can improve average confidence alignment, it does not necessarily produce a better ranking of examples by answerability. Neighbourhood Disagreement instead introduces an epistemic reliability signal from local weight stability, yielding higher target-accuracy coverage across datasets. Finally, we show that this signal is useful only when the frozen model's local neighbourhood contains meaningful task-relevant structure. The method does not create new task knowledge; it exposes uncertainty around an already partially competent model. Overall, Neighbourhood Disagreement provides a simple, deployment-friendly route toward safer surgical VQA: answer locally stable cases and defer uncertain ones to the surgeon.

References

1. Agrawal, A., Batra, D., Parikh, D., Kembhavi, A.: Don't just assume; look and answer: Overcoming priors for visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
2. Bai, L., Islam, M., Seenivasan, L., Ren, H.: Surgical-vqla: Transformer with gated vision-language embedding for visual question localized-answering in robotic surgery. arXiv preprint arXiv:2305.11692 (2023)
3. Bordes, F., Pang, R.Y., Ajay, A., Li, A.C., Bardes, A., Petryk, S., Mañas, O., Lin, Z., Mahmoud, A., Jayaraman, B., et al.: An introduction to vision-language modeling. arXiv preprint arXiv:2405.17247 (2024)
4. Byun, J.Y., Park, Y.J., Corbeil, J.P., Ben Abacha, A.: Overconfidence and calibration in medical vqa: Empirical findings and hallucination-aware mitigation. arXiv preprint arXiv:2604.02543 (2026)
5. Ding, D., Yao, T., Luo, R., Sun, X.: Visual question answering in robotic surgery: A comprehensive review. IEEE Access **13**, 9473–9484 (2025)
6. Fingerhut, A., Mayol, J.: Artificial intelligence for global surgery: Bridging gaps in low-resource settings (2026)
7. Gan, Y., Isola, P.: Neural thickets: Diverse task experts are dense around pretrained weights. arXiv preprint arXiv:2603.12228 (2026)
8. Geifman, Y., El-Yaniv, R.: Selective classification for deep neural networks. In: Advances in Neural Information Processing Systems (2017)
9. Geifman, Y., El-Yaniv, R.: Selectivenet: A deep neural network with an integrated reject option. In: Proceedings of the 36th International Conference on Machine Learning. pp. 2151–2159. PMLR (2019)
10. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
11. Hashemi, N., Mose, M., Østergaard, L.R., Bjerrum, F., Søgaard-Andersen, E., Fabrin, K., Tuckus, G., Friis, M.L., Rasmussen, S., Tolsgaard, M.G.: Closing the data gap: leveraging pretrained neural networks for robotic surgical assessment on limited clinical data. Journal of Robotic Surgery **20**(1), 39 (2025)
12. Li, X., Li, J., Li, F., Zhu, L., Yang, Y., Shen, H.T.: Generalizing vision-language models to novel domains: A comprehensive survey. ACM Transactions on Multimedia Computing, Communications and Applications (2025)
13. Manzano Rodriguez, A., Snoek, C.G., Schijven, M.P.: Bridging the gap: exposing the hidden challenges towards adoption of artificial intelligence in surgery. British Journal of Surgery **112**(11), znaf217 (2025)
14. Seenivasan, L., Islam, M., Krishna, A.K., Ren, H.: Surgical-vqa: Visual question answering in surgical scenes using transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 33–43. Springer (2022)
15. Varghese, C., Harrison, E.M., O'Grady, G., Topol, E.J.: Artificial intelligence in surgery. Nature medicine **30**(5), 1257–1268 (2024)
16. Wang, G., Bai, L., Nah, W.J., Wang, J., Zhang, Z., Chen, Z., Wu, J., Islam, M., Liu, H., Ren, H.: Surgical-lvlm: Learning to adapt large vision-language model for grounded visual question answering in robotic surgery. arXiv preprint arXiv:2405.10948 (2024)

17. Xiao, W., Gan, X., Gan, L.: Vl-calibration: Decoupled confidence calibration for large vision-language models reasoning. arXiv preprint arXiv:2604.09529 (2026)
18. Zhang, J., Huang, J., Jin, S., Lu, S.: Vision-language models for vision tasks: A survey. IEEE transactions on pattern analysis and machine intelligence **46**(8), 5625–5644 (2024)