Data Leakage in Kvasir-VQA-x1: Overlapping Images Threaten Evaluation Validity (Team Lama4Vision)

Mahdi Azmoodeh-Kalati[1], Mohammad Sadegh Maghareh[2], Saba Alavi[3] and Reza Lashgari[1,*]

[1]*Institute of Medical Science and Technology, Shahid Beheshti University, Tehran, Iran*
[2]*Computer Science Department, Amirkabir University of Technology, Tehran, Iran*
[3]*Artificial Intelligence Department, Shahid Beheshti University, Tehran, Iran*

Abstract

Kvasir-VQA-x1 is a large Gastrointestinal (GI) endoscopy visual question answering (VQA) dataset introduced for the MediaEval Medico 2025 challenge. It contains approximately 160,000 question–answer (Q&A) pairs grounded in about 6,500 endoscopic images.

In our analysis, we uncover a critical data leakage issue in this dataset: due to splitting at the Q&A pair level, the same images appear in both the training and test sets. Specifically, 3,821 image IDs (over half of all images) are present in both splits. This overlap violates the standard assumption that evaluation images remain unseen during training, thereby biasing performance metrics. As a result, models can effectively "remember" visual content from training images and answer test questions without true generalization, leading to inflated accuracy.

We qualitatively analyze the dataset structure and provide visual evidence of this leakage. We frame this as a serious threat to the validity of vision–language model evaluation and emphasize its implications for the reliability of benchmark results.

Finally, we propose a straightforward yet effective remedy: an *image-level split* that assigns all Q&A pairs associated with a given image to a single subset (train, validation, or test). This approach eliminates overlapping images between splits, ensuring a fairer evaluation of models' ability to generalize to new visual data. Through this work, we aim to raise awareness of the leakage issue and advocate for improved dataset-splitting practices in VQA and multimodal learning.

1. Introduction

Reliable evaluation is a cornerstone of machine learning, demanding that models be tested on data not seen during training[1]. Data leakage occurs when training and test sets share content, allowing models to cheat by memorizing rather than truly learning[2]. In computer vision and VQA, leakage often takes the form of overlapping images across splits. This undermines the integrity of results, since models can achieve artificially high performance by recalling training images during testing. Recent studies highlight that such overlap can grossly inflate metrics – in medical imaging tasks, improper splitting has boosted accuracy by up to 30% in some cases . Large vision-language models are especially susceptible: if they have inadvertently seen test images (or very similar ones) during training, their reported multimodal reasoning ability may be overstated[3]. Kvasir-VQA-x1 is a multimodal dataset for medical VQA in GI endoscopy, introduced in 2025 to benchmark reasoning-intensive VQA models. It extends the original Kvasir-VQA dataset (6,500 images, 59k QA pairs) with an order of magnitude more Q&A pairs (159k) generated via augmentation and large language models. The dataset has been adopted as

MediaEval'25: Multimedia Evaluation Workshop, October 25–26, 2025, Dublin, Ireland and Online

[*] Corresponding author.

[†] These authors contributed equally.

r_lashgari@sbu.ac.ir (R. Lashgari)

CEUR Workshop Proceedings (CEUR-WS.org)

the basis for the MediaEval Medico 2025 VQA challenge, aiming to drive advances in robust medical VQA Ensuring the evaluation is leakage-free is therefore crucial: an inflated benchmark could mislead researchers about model capabilities and progress. The Kvasir-VQA-x1 creators did attempt to separate training and test content. The official documentation states that the train split is for training only and test is held-out for final evaluation, with the intention that "no image or QA from the test set appears in the training set". However, we discovered that this intention was not met in practice due to the splitting strategy employed. In this paper, we analyze the Kvasir-VQA-x1 splits to quantify the overlap, discuss how it can bias results, and propose a remedy. We cite relevant studies on data leakage to underscore the implications for vision-language evaluation. Our contribution is to alert the community to this dataset issue and recommend an improved splitting approach to bolster the validity of MedVQA benchmarks.[4]

2. Task Description

2.1. Kvasir-VQA-x1 Dataset

Kvasir-VQA-x1 is a large-scale medical VQA dataset focused on GI endoscopy images. Each sample consists of an image and a complex question (with an answer) about that image, often decomposable into multiple sub-questions. The images (≈6.5k unique endoscopic images) were sourced from HyperKvasir and Kvasir-Instrument datasets and cover various GI findings (polyps, landmarks, etc.). Questions were semi-automatically generated and validated by clinicians to ensure realism and diversity (spanning yes/no, counting, localization, procedural, and reasoning types). The total QA pairs in Kvasir-VQA-x1 are 159,549, divided into predefined training and test splits. The training set contains ≈144k Q&A pairs and the test set ≈16k, as shown in the official HuggingFace dataset repository. The Kvasir-VQA-x1 dataset page on Hugging Face, showing distinct train (144k Q&A pairs) and test (16k Q&A pairs) splits. The dataset is part of the MediaEval Medico 2025 challenge. Each Q&A entry is associated with an img_id (image identifier). The creators split the data by Q&A pair, meaning each image-question pair was assigned to either the train or test set. This was done to ensure no identical question appears in both sets. In theory, such splitting would prevent trivial memorization of question-answer pairs. However, a consequence is that an image with multiple QAs can have some of its QAs in training and others in testing. The intention was to support generalization testing on novel questions, but it inadvertently introduced image overlap between splits. In other words, while exact Q&A pairs were not repeated, a large number of the same images appear in both train and test sets under different questions.

3. Analysis of Data Leakage

We analyzed the train/test splits and found a significant overlap at the image level. Out of approximately 6,500 total unique images in Kvasir-VQA-x1, 3,821 images are present in both the training and test sets. The training split contains Q&A pairs for 6,212 unique images, and the test split covers 4,058 unique images.

If the splits were truly disjoint by image, the total number of unique images would be the sum $(6,212 + 4,058 \approx 10,270)$. Instead, the actual number of distinct images is only about 6,500, confirming a large intersection of images between the splits.

Only a few hundred images are exclusive to the test set (roughly 237 images, by calculation), meaning that the vast majority of test images (around 94%) were already seen in training (albeit with different questions). This represents a textbook case of *intra-dataset leakage*, where evaluation data is not truly unseen. Overlap analysis of Kvasir-VQA-x1 splits, showing that 3,821 image IDs are common to both train and test sets (out of ≈6,500 total images). In the snippet above, we see the train split has 143,594 QAs across 6,212 images, and the test split 15,955 QAs across 4,058 images, with an overlap of 3,821 images (hence the sum of unique IDs exceeds the total images). This indicates that more than half of all images (≈59%) appear in both sets, a severe data leakage. Implications: A vision-language model trained on this data will have seen nearly all test images during training. Even though the specific questions in the test may differ, the model could rely on memorized visual features from those images. For example, if a training question asks "How many polyps are present in this image?" and a test question on the same image asks "Is there any bleeding in the image?", the model has already been exposed to the image's content and may more easily answer the test question (having recognized the scene in training). The evaluation no longer measures generalization to new visual data, but rather only to new questions on familiar images. This inflates performance metrics: models can achieve higher accuracy or BLEU scores than they would on truly novel images, since part of the task (interpreting the image) was solved during training. Our findings resonate with prior research showing that overlapping training and test data yields overly optimistic results. In medical image analysis, Tampu et al. demonstrated that improper splitting (e.g., slices from the same 3D scan in both train and test) exaggerated performance by 5–30%. Similarly, in VQA and multimodal benchmarks, Chen et al. observed that models sometimes answer visual questions correctly without even looking at the image if those images or questions were seen during pre-training. In our case, a model could potentially answer certain test questions using spurious cues learned from the train-phase exposure to that image (a form of unintended cheating).

From a validity standpoint, this leakage undermines the claims one can make about a model's capability. If a model achieves, say, 80% accuracy on the Kvasir-VQA-x1 test set, we cannot be sure how much of that success is due to genuine generalization versus simply recognizing images it has seen before. The integrity of the benchmark is compromised. This is especially problematic for the MediaEval challenge: teams might report high scores that do not translate to real-world performance on new patient cases, since the evaluation is inadvertently easier. In summary, the Q&A pair-level split in Kvasir-VQA-x1 creates a train-test contamination that biases results. We did not run new experiments due to time constraints, but based on the literature and the scale of overlap, it is reasonable to expect a noticeable inflation of evaluation metrics. This calls for an immediate reconsideration of the dataset split for any future use of Kvasir-VQA-x1 in benchmarking.

3.1. Proposed Solution: Image-Level Splitting

To resolve the leakage issue, we propose restructuring the dataset splits at the image level. The core idea is simple: ensure that no image appears in more than one split. This aligns with standard best practices, where medical imaging challenges commonly perform data partitioning by patient or study to avoid cross-contamination.

In the context of Kvasir-VQA-x1, splitting by image means that all Q&A pairs associated with a given image ID will belong to a single subset (train, validation, or test). The procedure can be summarized as follows:

1. **Collect Unique Images:** Compile the list of all unique `img_id` values in the dataset (approximately 6,500 IDs). These represent the distinct endoscopic images available.
2. **Random Split by Image:** Randomly divide the image IDs into new splits with no overlap. For example, 70% of the images can be used for training, 15% for validation, and 15% for testing (or another suitable ratio). This yields disjoint image sets. Stratification can

also be applied to maintain similar distributions of question types or complexities across splits.

3. **Assign Q&A Pairs:** For each Q&A pair in the dataset, assign it to the split corresponding to its image ID. All questions about a particular image will end up in the same split. This may slightly reduce the total number of Q&A pairs per split compared to the original setup, but ensures the integrity of the evaluation.

By following this method, the train set will contain images that are completely separate from those in the test set (and similarly for the validation set). A model trained on the new training split would never see even a single pixel of any test image during training. Consequently, performance on the test split will more accurately reflect the model's ability to generalize to new images, providing a far more realistic and rigorous evaluation for a VQA system.

Benefits. First, it eliminates data leakage—the evaluation will be fair by design, upholding the principle that no test sample (image) is seen during training. This prevents inflation of accuracy metrics; any improvement in performance should result from better visual reasoning or generalization, not memorization. Second, it provides a more challenging and meaningful benchmark. In the medical domain, deployed models must handle entirely new images from unseen patients. An image-level split simulates this scenario by holding out images that the model has never encountered, thus producing test results that better indicate real-world performance. Third, it encourages models to focus on learning general visual features and multi-modal alignments, since they cannot rely on recalling specific training images. In contrast, the current leakage could reward models that implicitly memorize training imagery. Lastly, introducing a separate validation split (if not already present) allows more principled model development—hyperparameters can be tuned on validation data, while the test set remains reserved for final evaluation, all with non-overlapping images.

We acknowledge that an image-level split might slightly alter the distribution of questions, since certain images have many associated Q&As. However, this is a minor concern compared to the benefits of preventing leakage. The overall dataset size remains sufficient for robust model training and evaluation (e.g., using 70% of $\sim$160k Q&As yields approximately 112k for training and 24k for testing if 15% of images are reserved for test). If needed, sampling strategies can maintain the original Q&A counts while enforcing unique image IDs per split. Finally, any augmented variants of an image should be treated as the same image for splitting purposes, to avoid the model seeing altered versions of test images during training.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT (OpenAI GPT-4) in order to: Improve wording and grammar and assist in structuring sections. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] I. E. Tampu, A. Eklund, N. Haj-Hosseini, Inflation of test accuracy due to data leakage in deep learning-based classification of oct images, Scientific Data 9 (2022) 580.

[2] P. Ramos, R. Ramos, N. Garcia, Data leakage in visual datasets, arXiv preprint arXiv:2508.17416 (2025).

[3] L. Chen, J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, J. Wang, Y. Qiao, D. Lin, et al., Are we on the right way for evaluating large vision-language models?, Advances in Neural Information Processing Systems 37 (2024) 27056–27087.

[4] S. Gautam, M. A. Riegler, P. Halvorsen, Kvasir-VQA-x1: A Multimodal Dataset for Medical Reasoning and Robust MedVQA in Gastrointestinal Endoscopy, arXiv (2025). doi:10.48550/arXiv.2506.09958. arXiv:2506.09958.