Cambridge Prisms: Water

www.cambridge.org/wat

Multi-class sewer defect detection with vision-language models

Riccardo Taormina and Job Augustijn van der Werf

Delft University of Technology, Netherlands

Research Article

Cite this article: Taormina R and van der Werf JA (2026). Multi-class sewer defect detection with vision-language models. *Cambridge Prisms: Water*, **4**, e12, 1–11
https://doi.org/10.1017/wat.2026.10019

Received: 06 October 2025
Revised: 26 December 2025
Accepted: 17 February 2026

Keywords:
asset management; sewer inspection; artificial intelligence; visual language models

Corresponding author:
Riccardo Taormina;
Email: r.taormina@tudelft.nl

Abstract

Proactive sewer asset management requires accurate condition assessment, yet CCTV inspections remain costly because interpretation is manual. We evaluated 18 vision-language models (VLMs) in a zero-shot setting for automated classification of six sewer defect types using a curated dataset. Each model produced a defect label, a short explanation, and a confidence score. OpenAI proprietary models outperformed open-source ones. GPT-4.1 mini achieved the highest macro-F1 score (0.50), outperforming much larger models, especially for surface damage and cracks/breaks. Some open-source models, such as LLaMa 4 (16x17B) and Qwen2.5-VL (32B), performed above random guessing but remained behind the proprietary models. All models failed to detect production errors, the most difficult class, and performed poorly on deformations. Confidence scores were generally unreliable, with little distinction between correct and incorrect predictions. Textual-output analysis showed that models sometimes described defects accurately even when the assigned label was wrong, although major hallucinations remained. We conclude that VLMs show some promise for sewer asset management, but they are not ready for deployment. Future work should focus on adding asset metadata to prompts and fine-tuning open-source models, especially since larger, newer, and more expensive OpenAI models did not outperform smaller ones, although confirmation requires a more thorough statistical analysis.

Impact statement

Sewer failures impose high economic, social and environmental costs: emergency excavations disrupt traffic, contaminate waterways and accelerate infrastructure ageing. Current AI solutions can reduce manual workload but require millions of labelled images, a barrier for most utilities. Our study shows that modern *vision-language models* (VLMs) can perform **multi-class sewer defect detection *zero-shot***, i.e., without domain-specific training, and with usable accuracy. Furthermore, smaller open-source models already achieve performance above random baselines and could be fine-tuned with curated datasets to provide practical, cost-effective tools for sewer defect detection. Key impacts are:

- **Reduced dependence on labelled data.** Utilities can deploy AI models that require fewer curated labels, enabling a more efficient approach to sewer inspection.
- **Democratisation of AI tools.** Open-source VLMs show promising performance and, with fine-tuning, could provide affordable solutions for municipalities with limited budgets.
- **Path to more explainable asset management.** If aligned, natural-language rationales generated by VLMs may foster trust and facilitate human – AI collaboration, a prerequisite for regulatory acceptance of automated inspections.

These advances suggest a potential pathway toward scalable, affordable AI tools for sewer management, supporting utilities in moving from costly manual inspections to proactive infrastructure maintenance.

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Introduction

Urban drainage networks form a critical yet largely invisible infrastructure that protects public health and the environment. Large cities manage thousands of kilometres of sewer pipes whose combined replacement value is measured in billions of euros. Localised failures, such as collapses, severe root intrusions, or local settling, can trigger emergency repairs, traffic disruption and pollution incidents. To avoid unexpected Routine condition assessment is therefore indispensable, and closed-circuit television (CCTV) surveys remain the sector's work-horse. However, frame-by-frame human review is labour-intensive, time-consuming and prone to subjectivity (Dirksen et al., 2013; Tscheikner-Gratl et al., 2019).

Early automation attempts relied on handcrafted image processing (edge detectors, morphological filters), but the past decade has seen a shift toward convolutional neural networks (CNNs) and other deep learning-based computer vision models (Haurum and Moeslund, 2020; Haurum et al., 2022a, 2022b). State-of-the-art pipelines achieve high accuracy for binary and multi-class

sewer defect classification (Xie et al., 2019; Haurum and Moeslund, 2021). Nevertheless, these deep learning systems face two persistent barriers: (i) they are data-hungry, requiring thousands to millions of carefully labelled images for supervised training, (ii) they remain opaque, offering only limited insight into why a decision was made, thereby limiting operator trust. Recent efforts show that transfer learning and self-supervised learning can reduce the annotation burden (Yildizli et al., 2025). However, these approaches still require substantial amounts of labelled data for downstream fine-tuning and do not address the fundamental lack of interpretability inherent to deep learning computer vision architectures. More interpretable CNNs are being developed (George et al., 2025), though the interpretation is still significantly complex and not straightforward.

An emerging class of models may offer a way forward. Large vision-language models (VLMs) – a subclass of multimodal large language models – jointly embed images and text so that a single prompt can query both modalities (Zhang et al., 2024). Thanks to their broad training and cross-modal alignment, VLMs can generate predictions without task-specific fine-tuning and provide human-readable rationales for their decisions. When additional textual context (e.g., defect definitions, maintenance manuals, or sensor readings) is supplied, a VLM can leverage this external knowledge and deliver zero-shot predictions: generate outputs for tasks or images it has never been explicitly trained on. For sewer inspection, the approach offers three main benefits:

1. **Context fusion**: domain knowledge from heterogeneous data streams can be injected at inference time, allowing the model to reason beyond the raw pixels.
2. **Interpretability**: Interpretability: prompts can request concise natural-language rationales that summarise salient visual and contextual cues linked to each prediction. While these explanations are post hoc and not necessarily faithful representations of the model's internal reasoning, they can assist inspectors in understanding model outputs and identifying potential failure modes.
3. **Interactive workflows**: inspectors may converse with the model, refine prompts, or correct misclassifications in real time by interacting with the AI model; these interactions can be recorded for further model fine-tuning.

In a first proof-of-concept, Taormina and van der Werf (2024) prompted VLMs for *binary* sewer defect detection and compared the results against a state-of-the-art CNN developed using more than one million SewerML images (Xie et al., 2019; Haurum and Moeslund, 2021). On a balanced 200-image subset, the best VLM (OpenAI GPT-4V) achieved a zero-shot F_1 of **0.65** versus **0.81** for the CNN. A subsequent cross-analysis (overlaying the binary predictions onto the ground-truth multi-class labels) showed GPT-4V already matching, and for some high-impact classes even surpassing, the CNN. However, this apparent per-class performance emerged only *indirectly*: the VLMs were never asked to identify which type of defect was present, only whether defect existed.

The analysis could therefore conflate correct detection with coincidental overlap: a frame labelled as *Roots* in the ground truth might be flagged as defective for unrelated reasons (e.g., a visible crack), and this would still count as a "true positive" under binary evaluation. Whether the VLMs' predictive skill generalises to actual multi-class classification has yet to be determined.

Additionally, the VLM landscape has evolved rapidly since the pilot study. Proprietary models now include native multimodal architectures and enhanced visual reasoning via chain-of-thought prompting. A growing ecosystem of open-source VLMs has emerged, ranging from compact architectures to larger-scale foundation models. Some of the smallest models can, in principle, be fine-tuned for specific tasks on consumer-grade GPUs, yet systematic benchmarks evaluating their zero-shot performance remain unavailable. Moreover, there is no prior study that quantifies how detection performance varies with key inference parameters such as model size, decoding temperature (which adjusts how confidently or cautiously the model makes predictions) and chain-of-thought reasoning efforts.

To close these gaps, we conduct the first systematic evaluation of multi-class zero-shot sewer defect detection using a broad spectrum of VLM architectures. Specifically:

1. We evaluate a total of 18 VLMs, including six proprietary OpenAI vision models and 12 open-source models, including the Gemma 3, Qwen2.5-VL and Llama families. We also include very compact models (i.e., < 5 B parameters) that may be suitable for fine-tuning in resource-constrained settings.
2. We compare their classification performance on a curated subset of 100 SewerML images, annotated with five common defect types, plus a no-defect class.
3. We analyse how key inference factors such as model size, decoding temperature and reasoning efforts influence accuracy, offering insights into practical deployment trade-offs.
4. We assess hallucination tendencies by comparing each model's stated confidence with the correctness of its predictions, quantifying reliability alongside accuracy.
5. We qualitatively assess the level of hallucination by manually comparing the model's textual outputs with the images to further evaluate the reliability of the VLM outputs.

Methodology

Dataset

We utilise a balanced subset of 100 images from the SewerML benchmark dataset (Haurum and Moeslund, 2021), selecting six classes: Cracks/Breaks (*RB*), Surface Damage (*OB*), Production Errors (*PF*), Roots (*RO*), Deformations (*DE*) and No Defect (*ND*). The resulting split contains 10 images per defect class and 50 no-defect cases. This selection balances criticality (e.g. root intrusion and breakage are urgent intervention triggers), prevalence (conditions occurring more often in real-world occurrences) and perceptual complexity (deformations, production errors and surface damage are often subtle and ambiguous). The SewerML dataset is released for non-commercial research use only, and to the best of our knowledge, is therefore unlikely to have been included in the training data of commercial VLMs evaluated in this study.

Vision-language models

Fusion architectures

VLMs can be broadly subdivided based on how they integrate visual and textual information (Wang et al., 2022). In *early-fusion* architectures, image and text tokens are processed jointly from the very first Transformer layer. This is typically achieved by converting image patches into discrete visual tokens – often via a lightweight vision encoder – and concatenating them directly with text tokens. The resulting token sequence is then passed through a single multimodal Transformer. Early-fusion models are considered

https://doi.org/10.1017/wat.2026.10019 Published online by Cambridge University Press

"natively multimodal" because they are trained end-to-end across modalities, enabling tightly coupled cross-modal representations. However, this approach requires large-scale co-training on image-text pairs and is computationally intensive.

In contrast, *late-fusion* models adopt a two-stream design: a dedicated vision encoder first processes the image into a compact feature representation, which is then passed to a pre-trained language model via cross-attention or projection layers. This modularity allows one to leverage existing language models without retraining them from scratch. Many late-fusion VLMs rely on vision encoders such as CLIP (Contrastive Language-Image Pre-training, Radford et al., 2021), which aligns image and text embeddings using a contrastive loss and is typically paired with a Vision Transformer (ViT) backbone (Dosovitskiy et al., 2020). Others use variants like SigLIP (Sigmoid Loss for Language Image Pre-Training), which replaces the contrastive softmax with a sigmoid loss for improved training stability, particularly in smaller or resource-constrained models (Zhai et al., 2023). While late-fusion architectures are generally more flexible and efficient to fine-tune, they may suffer from weaker cross-modal coupling compared to early-fusion approaches that learn joint representations from the start.

Model paradigms

Instruction-following models are trained to produce direct answers in natural language based on user prompts, prioritising speed and linguistic alignment. **Reasoning** models (e.g. OpenAI's o-series) generate internal chain-of-thought steps (Wei et al., 2022) before emitting an answer, often integrating multi-step reasoning or tool-use modules to reach conclusions more reliably.

Selected models

We evaluate 18 VLMs, covering both proprietary and open-source families, with varying design philosophies and parameter scales. Our selection spans early-fusion models (e.g. GPT-4o, Llama 4) and late-fusion architectures (e.g. Qwen, Mistral and Gemma). It also includes reasoning models (o3 and o4-mini) to complement the vast majority of architectures, which are tuned for instruction following. For most families, we include multiple variants of different sizes in order to explore how scale affects performance under consistent zero-shot prompting.

Proprietary models (OpenAI)

- **GPT-4o and GPT-4o-mini**: Early-fusion "Omni" models trained natively on text, images and audio within a single unified Transformer (Hurst et al., 2024). The mini variant is a distilled version optimised for speed and resource efficiency, with trade-offs in accuracy.
- **GPT-4.1 and GPT-4.1-mini**: Successors to the aforementioned Omni models, but without the audio modality (OpenAI, 2024a). The models maintain early-fusion vision-text integration and outperform GPT-4o on complex multimodal benchmarks. The mini version preserves core capabilities while reducing inference cost.
- **o3 and o4-mini**: Leading members of OpenAI o-series of *reasoning* systems that generate rich internal chain-of-thought (OpenAI, 2024b). As of our study date (until July 2025), o3 achieves OpenAI's strongest vision-reasoning scores to date, while the o4-mini variant retains superior vision capabilities with reduced computational overhead.

Open-source models

- **Gemma 3 (1B, 4B, 12B, 27B)**: Google's open-source VLMs is a late-fusion decoder-only Transformer that ingests vision embeddings produced by a frozen SigLIP encoder; the image tokens are concatenated directly to the text sequence (Team Gemma et al., 2025).
- **Qwen 2.5-VL (3B, 7B, 32B, 72B)**: Combines a dynamic-resolution ViT vision encoder (i.e., window attention) with a Qwen 2.5 decoder, joined through a multi-layer perceptron that compresses patch features into text-sized embeddings for late fusion (Bai et al., 2025). Trained on over 4 trillions multimodal tokens, it excels at document parsing, fine-grained grounding (bounding-boxes/points) and long-video understanding.
- **Llama 3.2 Vision (90B)**: Hybrid-fusion VLM: a ViT-based encoder feeds embeddings through cross-attention adapters interleaved across the Llama 3.1 decoder, achieving strong visual reasoning, captioning and grounding (Meta AI, 2024). Llama 4 Scout (16 ×17B, 109B total parameters): Multimodal Mixture-of-Experts model that activates a single 17B expert per token; early-fusion text-vision pre-training and expert routing deliver (AI, 2025).
- **Mistral Small 3.2 (24B)**: Update of Mistral Small 3.1 (Mistral AI, 2025). Keeps the 24B latency-optimised architecture (lightweight vision encoder feeding a shallow decoder), but adds fresh training data and instruction-tuning refinements for higher accuracy.
- **Granite-3.2 Vision (2B)**: Document-focused VLM developed by IBM. Late-fusion design: a frozen ViT-based vision encoder feeds a small projector whose outputs are concatenated to the prompt of a 2B-parameter decoder-only (Granite Vision Team et al., 2025).
- **Moondream 2 (1.8B)**: Late-fusion VLM that pipes a frozen SigLIP ViT encoder into a Phi-1.5 language decoder via a lightweight projector (Moondream AI, 2024). Aimed at mobile or in-browser inference.

Pipeline development and prompting

We used the OpenAI Python client v1.86.0 for OpenAI models, and Ollama v0.11.6 with the Ollama Python client v0.5.3 for open-weights models. The main prompt instructed models to act as expert sewer-inspection analysts, assigning one SewerML defect class to each CCTV frame based on visual cues and definitions, defaulting to ND if no defect was present. The expected output was a JSON object with class, confidence (real number between 0 and 1) and a short explanation. While reliable for most models, some open-weight models (e.g., Gemma 4B, Llama 3.2, Llama 4, Moondream, Granite) occasionally or systematically produced malformed outputs. We therefore applied a lightweight post-processing step with GPT-4.1-mini, using a repair prompt to extract valid JSON or fix it with minimal modifications with respect to the original output.

Performance evaluation

To quantify the performance of the VLMs, we calculated the performance of the 18 VLMs using per-class F1 scores and the overall macro-F1 as the main metric, computed as the arithmetic mean of the F1-score for each of the individual defect classes (i.e., averaged F1 methodology as per Opitz and Burst, 2019). We selected the macro-F1 score as our main metric because, at this stage, we aim to evaluate how VLMs perform independently of class

https://doi.org/10.1017/wat.2026.10019 Published online by Cambridge University Press

frequencies observed in the real world, giving equal importance to all defect classes and preventing non-defects, typically easier to identify, from dominating the evaluation. We used this metric to rank the VLMs and investigate the influence of model size, temperature and reasoning efforts. Additionally, we report macro-averaged *precision* and *recall* to further characterise model behaviour, in particular to assess whether VLMs tend to favour false positives or false negatives across defect classes. We also evaluated whether the confidence scores provided by the models were meaningful by comparing their distributions for correct and incorrect predictions. Specifically, we tested whether lower confidence was associated with a higher error rate using the non-parametric Kolmogorov – Smirnov test.

Furthermore, we qualitatively examined the textual outputs of the models. This expert-driven evaluation aimed to detect hallucinations and assess whether the VLMs' reasoning aligned with the ground-truth defect class by comparing the explanatory text directly against the reference label rather than the predicted label alone. Within this assessment, we classified the combined label and reasoning output of the VLMs as Correct (C, when the label was correctly predicted), Understandable or Arguable Difference (*UAD*, when a label was incorrectly predicted but the reasoning was convincing when compared to the image), Disconnection between Reasoning and Label (*DRL*, when the text suggest a different label compared to the one given), Minor Hallucination (*MH*, when the reasoning was flawed but within the context of the image) and Complete Hallucination (*CH*, when there was no alignment between the image and the reasoning whatsoever). This evaluation adds a level of subjectivity, though this is also present in the labelled dataset (Dirksen et al., 2013). We recalculated the macro F1 score, treating both *UAD* and *DRL* as valid (i.e., correct) labels for evaluation purposes.

Results and discussion

Quantitative model performance

In total, 54 model-parameter combinations were evaluated, corresponding to 18 models tested across three temperature settings (0, 0.5, or 1) or, for reasoning models, three levels of reasoning effort (low, medium, or high). Figure 1 shows the macro-F1 scores across parameter settings for open-weight models (a), non-reasoning OpenAI models (b) and reasoning OpenAI models (c). The GPT-4.1 mini models performed the best consistently, ranking first, second and third overall with scores of 0.50 (temperature = 0), 0.49 (temperature = 0.5) and 0.48 (temperature = 1). These were closely followed by the o3 model (0.47; reasoning = low) and the 4o model (0.46; temperature = 0). A closer comparison of the top-performing models reveals that GPT-4.1 mini achieves its leading macro-F1 score primarily due to its higher macro recall (0.547), indicating a stronger ability to detect defects across classes. This comes at the cost of a comparatively lower macro precision (0.476). By contrast, the best-performing o3 configuration exhibits a more balanced behaviour, with a lower macro recall (0.487) and a similar macro precision (0.460), resulting in a slightly lower overall ranking. GPT-4o ranks lower overall despite achieving a relatively high macro recall (0.513), as its substantially lower macro precision (0.424) leads to a higher rate of false positives, ultimately reducing its macro-F1 score. These results indicate that the superior performance of GPT-4.1 mini is largely driven by recall, whereas other high-performing models trade recall for more conservative predictions. A closer comparison of the top-performing models shows that GPT-4.1 mini attains the highest macro-F1 score primarily due to its stronger sensitivity to defects across classes (recall = 0.547, precision = 0.476). The best-performing o3 configuration exhibits a more balanced profile, with lower macro recall (0.487) and similar macro precision (0.460). GPT-4o achieves a higher macro recall (0.513), but its substantially lower macro precision (0.424) leads to more false positives and a reduced overall performance.

The highest-performing open-source model was Llama 4 (temperature = 1), with a macro-F1 score of 0.38, although its performance remained well below that of the best Open-AI models. Notably, all of the best-performing models are early-fusion VLMs, which may contribute to their superior performance by enabling tighter integration of visual and textual information. At the same time, Llama 4 is the largest open-weight model in our benchmark, and the sizes of the OpenAI models are undisclosed, leaving open whether performance improvements are due primarily to architecture or to scale. All of these models outperformed the baseline derived from 500,000 random-guess simulations (Figure 2), in which predictions were sampled uniformly across the six defect classes. Notably, only nine model combinations scored below the mean expected macro-F1 of random guessing (0.144), indicating that even the simpler open-source models generally provide predictive power beyond chance.

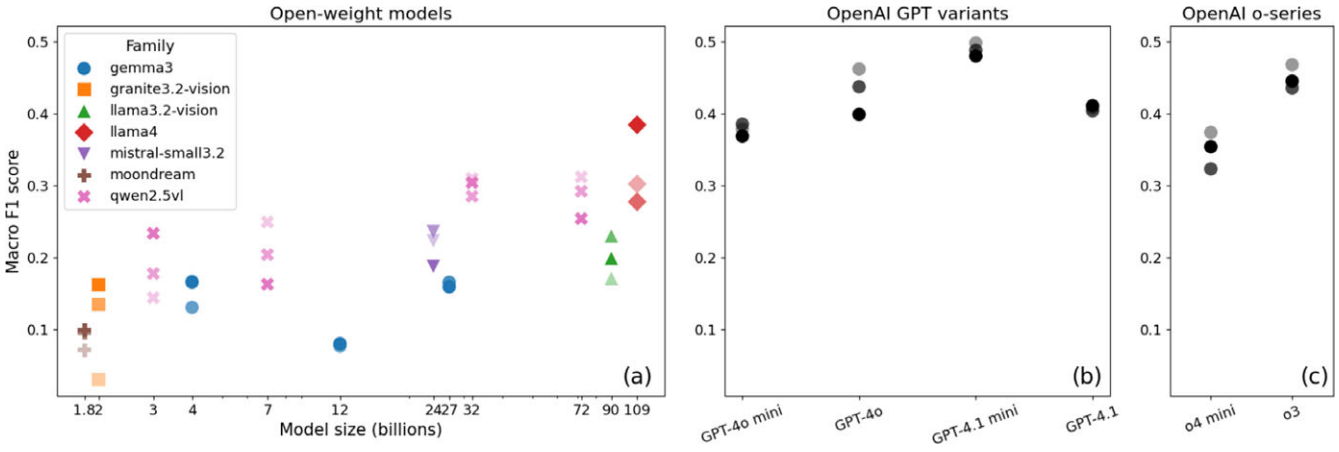


Figure 1. Macro-F1 scores across model families and parameter settings for open-weight models (a), OpenAI GPT variants without reasoning (b) and o-series models with reasoning (c). Identical colour/marker combinations denote results for the same model families across different settings. Transparency encodes either the temperature or the reasoning effort: the most transparent points correspond to temperature 0.0 or low reasoning effort, while the least transparent correspond to temperature 1.0 or high reasoning effort.

https://doi.org/10.1017/wat.2026.10019 Published online by Cambridge University Press

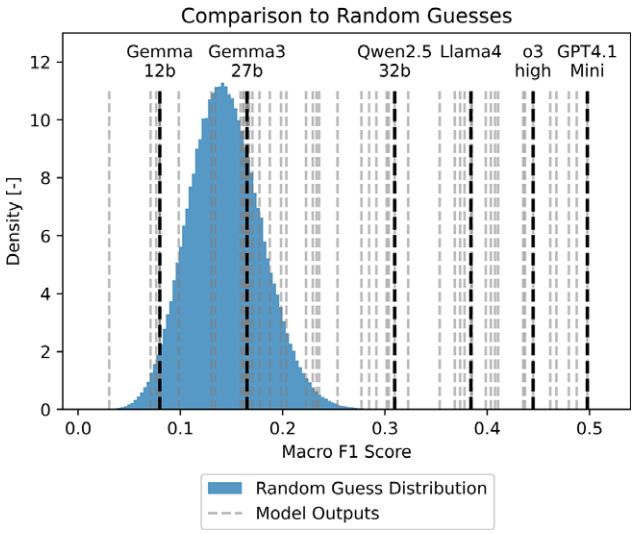


Figure 2. Model performance benchmarked against a population of 500,000 random guess models.

Table 1 details the performance of the best-performing parameter selection for each model type across individual defect classes. GPT-4.1 mini achieves particularly strong results for *RB* (0.74) and *OB* (0.57), which contributes directly to its superior macro-F1 score. The o3 model performs best on *ND* (0.83) and *RO* (0.74); its strong *ND* detection accounts for its higher weighted-F1 score, where contributions are weighted by class frequency. GPT-4o also performs competitively, achieving the top scores for *RO* (0.75) and *DE* (0.38). Importantly, none of the evaluated models successfully identified instances of the *PF* class, confirming it as the most challenging defect type, followed by *DE*. Among the open-source models, results are consistently lower than for the OpenAI models. Llama 4 and the Qwen family are relatively strong in detecting *ND*, but their performance across the remaining classes is generally below 0.5, highlighting substantial gaps in balanced defect detection.

Dependence of performance on model size

When examining the relationship between model size and performance for the open-source models (Figure 1), we observe that, in general, larger models tend to achieve higher macro-F1 scores, forming an approximate Pareto front. In particular, we see a steep improvement when moving from very small to moderately sized models, followed by a more gradual increase. There are, however, notable exceptions. For instance, Llama 3.2, despite being substantially larger than most open-source models, scores below much smaller models. Within the Qwen2.5-VL family, the best performance is achieved by the 32B model, after which the gains plateau, with the 72B model providing no further improvement. By contrast, within the Gemma 3 family, the 4b model performs on par with the larger 27b model; however, the Gemma models overall show relatively weak performance, especially when compared to the Qwen family.

The analysis of OpenAI models is more complex, as neither the number of parameters nor the architectural differences between model variants are fully disclosed. Nevertheless, our results indicate that GPT-4.1 mini achieves the highest macro F1 score overall, outperforming larger models, including o3, which OpenAI describes as its most capable model for visual reasoning. Despite comparable claims, o4-mini was the weakest OpenAI model we tested, scoring

Table 1. Performance of the best-performing model combination (GPT-4.1 mini, temperature = 0)

Model	Parameter	macro F1 score	weighted F1 score	RB	ND	RO	OB	PF	DE
gpt_4.1_mini	temperature = 0.0	**0.50**	0.58	**0.74**	0.71	0.60	**0.57**	0.00	0.37
o3	reasoning = low	0.47	**0.61**	0.58	**0.83**	0.74	0.42	0.00	0.24
gpt_4o	temperature = 0.0	0.46	0.59	0.67	0.77	**0.75**	0.20	0.00	**0.38**
gpt_4.1	temperature = 1.0	0.41	0.54	0.60	0.73	0.67	0.27	0.00	0.19
llama4_16x17b	temperature = 1.0	0.38	0.55	0.40	0.79	0.56	0.42	0.00	0.14
gpt_4o_mini	temperature = 0.5	0.38	0.52	0.53	0.72	0.67	0.39	0.00	0.00
o4_mini	reasoning = low	0.37	0.52	0.38	0.75	0.62	0.32	0.00	0.17
qwen2.5vl_72b	temperature = 0.0	0.31	0.42	0.19	0.58	0.59	0.23	0.00	0.29
qwen2.5vl_32b	temperature = 0.0	0.31	0.50	0.30	0.79	0.46	0.31	0.00	0.00
qwen2.5vl_7b	temperature = 0.0	0.25	0.44	0.00	0.73	0.46	0.30	0.00	0.00
mistral-small3.2_24b	temperature = 0.5	0.24	0.34	0.20	0.49	0.44	0.14	0.00	0.13
llama3.2-vision_90b	temperature = 0.5	0.23	0.40	0.00	0.65	0.35	0.24	0.00	0.13
qwen2.5vl_3b	temperature = 1.0	0.23	0.41	0.00	0.68	0.38	0.20	0.00	0.13
gemma3_27b	temperature = 0.5	0.17	0.21	0.20	0.27	0.33	0.19	0.00	0.00
gemma3_4b	temperature = 0.0	0.17	0.32	0.17	0.56	0.00	0.27	0.00	0.00
granite3.2-vision_2b	temperature = 1.0	0.16	0.21	0.19	0.29	0.11	0.18	0.00	0.20
moondream_1.8b	temperature = 1.0	0.10	0.25	0.00	0.47	0.12	0.00	0.00	0.00
gemma3_12b	temperature = 0.0	0.08	0.05	0.00	0.00	0.31	0.17	0.00	0.00

Note: Scores are reported as overall macro-F1 and weighted-F1, along with class-specific F1 scores for each defect type: RB (cracks, breaks, collapses), ND (no defect), RO (roots), OB (surface damage), PF (production errors) and DE (deformations).

https://doi.org/10.1017/wat.2026.10019 Published online by Cambridge University Press

well below GPT-4.1 across our benchmark. These findings suggest that, at least for the task considered, OpenAI model size or flagship status does not directly translate into superior performance for sewer defect assessment.

Role of temperature and reasoning effort

Temperature controls the determinism of the predictions, while reasoning effort regulates the computational resources allocated to generating a response. For the temperature-based (non-reasoning) models, we observe no consistent effect of temperature among the open-weight models. Even within the same family (e.g., Qwen2.5), different model sizes sometimes perform better at higher temperatures and sometimes at lower ones. The best-performing open-source model, Llama 4, achieved its highest score at the highest temperature setting (Figure 1a). By contrast, the OpenAI GPT variants, and in particular GPT-4.1 mini, consistently ranked among the top performers across all settings (Figure 1b). While this model performed well regardless of temperature, the GPT variants as a whole tended to achieve better results at lower temperatures, which is consistent with the expectation that higher determinism yields more reliable predictions. However, the role of temperature in shaping performance remains unclear. Regarding reasoning effort, both o3 and o4-mini attained their best results with low reasoning effort (Figure 1c), a counterintuitive finding given that higher reasoning levels are expected to improve performance. It is important to note that these results are based on single runs per configuration, and additional experiments would be required to establish their statistical robustness.

Analysis of confidence in the predictions

To aid interpretation of the VLM outputs, we also requested a confidence score (0–1) alongside each prediction. Most models largely hallucinate these values, reporting consistently high confidence (≥ 0.8) even for incorrect classifications (Figure 3). The Kolmogorov – Smirnov tests showed no significant difference ($p \leq 0.05$) between the confidence distributions for correct and incorrect predictions in the majority (77.3%) of the models. Among the remainder, most were significantly more confident when correct, while one model (Moondream 1.8b) reported higher confidence for its misclassified outputs.

Despite some statistically significant differences, the reported confidence values for the open-weight models are not practically meaningful. For example, Qwen2.5-VL (72b, temperature 0) shows a significant difference, but confidence remains consistently high (> 0.8) for both correct and incorrect predictions (Figure 3). Among the OpenAI models, o3 stands out: while higher reasoning effort does not yield higher macro-F1 scores, it does produce more informative confidence values. In the high reasoning setting, the confidence for incorrect predictions is clearly lower than that for correct ones, suggesting that increased reasoning may help calibrate confidence even if it does not improve classification accuracy.

Qualitative error analysis

To extend the analysis beyond macro-level performance indicators, we also examined the 5,400 combined image-text outputs produced by the models. Certain confusion pairs reoccurred systematically across VLMs. Harmless discolouration in plastic or re-lined pipes (often due to biofilm or thin films) (ND) was frequently misclassified as *surface damage* (OD), despite being commonplace and non-problematic. Models sometimes flagged stronger discolouration at the lower pipe wall as defects, but inconsistently treated similar cases as non-defective. Nonetheless, this led to the major occurrence of errors with pairs where *ND* was mislabelled as *OB* (515 instances across all model combinations). Production errors (PF) – typically wrinkling in plastic or re-lined walls – were often misinterpreted as *deformations* (DE), with textual descriptions referring to changes in pipe shape. This tendency was present across all VLMs, including higher-performing models, indicating a systematic difficulty in distinguishing production errors from deformations.

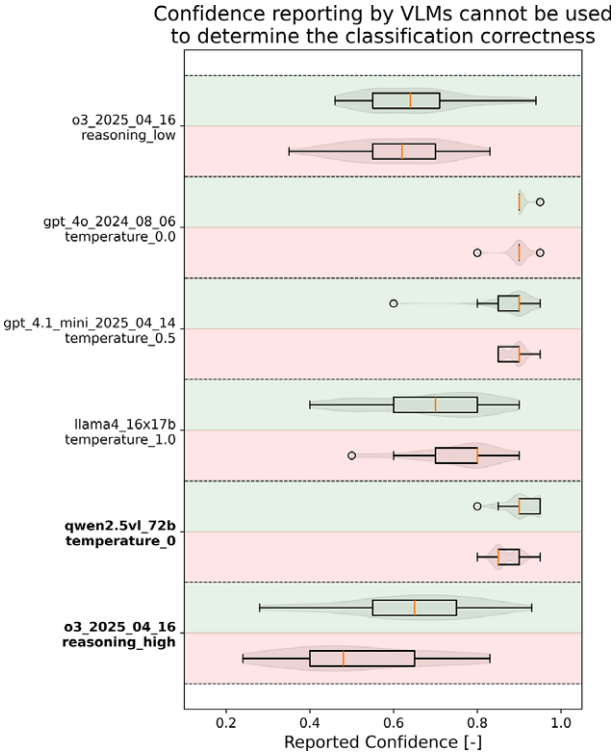


Figure 3. Distributions of the level of confidence reported by the top four models, the highest statistically significant difference (Qwen2.5) and OpenAI's o3 with high reasoning. The shaded green (top row for each model) shows the correctly labelled distribution, and red (bottom row for each model) the wrongly labelled data. Boxplots layouts follow the original Tukey's definition. Bold label on the y-axis indicates a statistically significant (KS test, $p \leq 0.05$) difference.

Table 2 shows the distribution of qualitative labels across all model combinations, which were assigned to predictions to capture not only correctness but also how well the reasoning aligned with the image across the different classes. From the analysis of the textual outputs, only seven predictions (0.13%) were classified as *DRL*, i.e., cases where the reasoning correctly described the image but the label was incorrect. This confirms that label-only hallucination is

Table 2. Overview of image label and prevalence of the qualitative labels (calculated as % per image label type)

	RB (%)	ND (%)	RO (%)	OB (%)	PF (%)	DE (%)
Total CH	49.4	40.4	37.7	41.1	83.0	80.4
Total MH	10.2	3.1	21.3	6.2	2.8	6.2
Total DRL	0.6	0.2	0.0	0.0	0.0	0.0
Total C	39.4	56.3	40.6	40.4	0.4	10.9
Total UAD	0.4	0.0	0.4	12.3	13.8	2.5

https://doi.org/10.1017/wat.2026.10019 Published online by Cambridge University Press

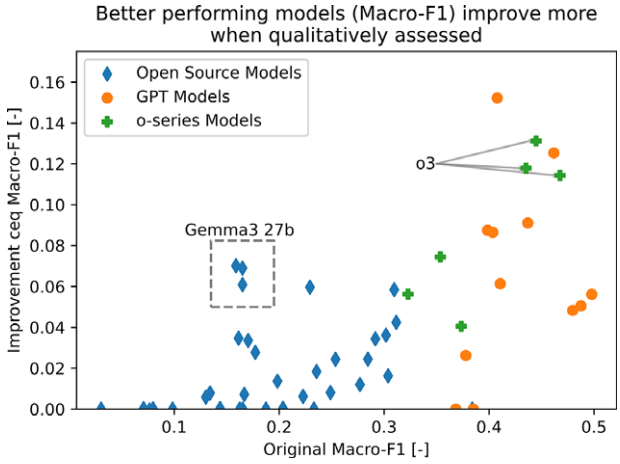


Figure 4. The improvement between the F1 score based on the labels alone and that computed including the qualitative assessment of the textual output. Gemma 3 (27b) model and the o3 models are highlighted: o3 shows no sensitivity to reasoning level and Gemma 3 (27b) shows consistent improvement with no clear sensitivity to temperature.

not a major issue. By contrast, a larger share of outputs (327; 6.2%) fell into the *UAD* category, where the predicted label was wrong but the reasoning could be considered plausible given the image. For the reasoning models, the level of reasoning had little effect on the number of *UADs*: o3 produced 8, 8 and 9 instances for low, medium and high reasoning, respectively, while o4-mini produced 4, 3 and 4. After recalculating the macro-F1 score by treating both *UAD* and *DRL* as correct labels, higher-performing proprietary models showed the greatest relative improvement, as shown in Figure 4. A notable exception was Gemma 3 (27b), which achieved the highest relative improvement ratio (0.44), largely due to its weaker baseline performance. Model size was not found to be correlated with improvements in macro-F1 score. The inclusion of the textual output alters the ranking of the top-performing models. GPT-4o (temperature = 0) and o3 (low and high reasoning) surpass the previously best-performing GPT-4.1 mini. Among the open-source models, Llama 4 (temperature = 1) remained the highest ranked. None of the models that originally performed worse than random guessing improved beyond the mean random baseline.

Table 2 also shows that the rate of *CH* was approximately twice as high for Production Error (PF) and Deformation (DE) compared to the other labels, with these two classes also having the lowest frequency of correctly classified (*C*) instances. These insights mirror the per-class results detailed in Table 1 for all models.

Visual analysis of prediction-explanation consistency

To further contextualise the qualitative error analysis, Table 3 presents representative CCTV images together with the predicted labels and textual explanations generated by three selected VLMs: GPT-4.1 mini (temperature = 0.0), o3 (reasoning = low) and Qwen-2.5-32B (temperature = 0.0). These models were selected to span the highest-performing proprietary systems and one of the best open-source alternatives, given their reduced size. The first row shows that severe structural failures, such as collapses and breaks, are detected reliably across all models, with both predicted labels and explanations aligned with the visual evidence. In contrast, more ambiguous classes exhibit substantial divergence. For instance, in the *OB* example, Qwen-2.5-32B predicts an *RB* label, constituting a *UAD* case. The explanation refers to a break in the pipe wall, which could be plausible given the severity of the surface damage on the left side of the image. The *Production Error* (PF) example illustrates the recurring confusion pattern of mislabelling this class as *Deformation* (DE), reflecting the inherent difficulty in visually distinguishing between these classes.

The *Root Intrusion* (RO) example is also informative, as the hair-like protruding strands are genuinely difficult to observe, even for human inspectors. While o3 correctly predicts RO, the most prominent root-like features are located slightly counterclockwise (approximately between the 8 and 10 o'clock positions), suggesting that this correct prediction may reflect a borderline or favourable case rather than precise localisation. By contrast, the prediction produced by Qwen-2.5-32B constitutes a minor hallucination (*MH*): although the assigned label is incorrect, the explanation refers to visible surface conditions that overlap with characteristics of the *surface damage* (OB) class. Finally, the *deformation* (DE) illustrates that while GPT-4.1 mini provides an explanation that is coherent and well aligned with the image content, the remaining models generate descriptions that are completely disconnected from the visual evidence, representing clear cases of hallucinated reasoning (CH).

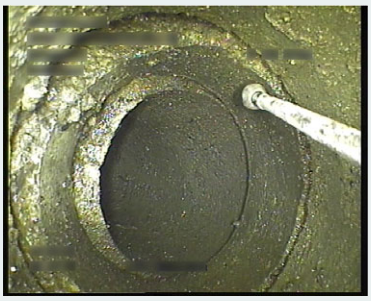
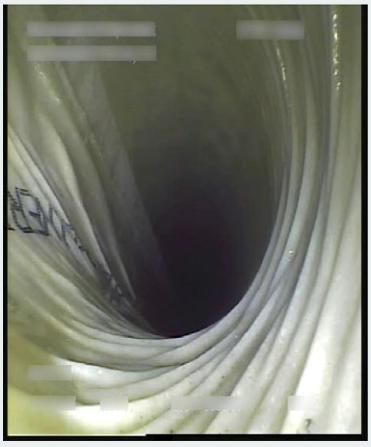
Discussion

Our results demonstrate the potential of VLMs to support the visual inspection of sewer systems. Proprietary models consistently outperformed open-source ones in a zero-shot setting, yet some open-source models (i.e., LLama 4 or Qwen2.5-VL) showed encouraging results and could be further improved through fine-tuning. While the confidence levels reported by the models were generally unreliable, the performance of the o3 model suggests that aligning prediction confidence with reasoning ability may provide a pathway towards better-calibrated outputs. A promising direction would therefore be to fine-tune open-source models on recreated datasets that explicitly couple prediction, reasoning and calibrated confidence. Such efforts could strengthen both interpretability and trustworthiness while maintaining the advantages of openly available VLMs. All models failed to detect Production Errors (PF) and performed poorly on Deformations (DE). For both defects, incorporating contextual information appears important to reduce hallucinations and improve accuracy. In the case of deformations, model outputs often referred to the perceived oval shape of pipes instead of the expected round shapes. Similarly, production errors in relined sewer pipes, such as material folding, were frequently misinterpreted as deformations. Providing contextual cues in the prompts. This could be done by linking georeferenced asset databases in an automated setting, and could therefore improve the reliability of model outputs.

An important limitation is that our analysis does not incorporate the additional complexity of reporting standards such as EN13508-2 (CEN, 2003), since SewerML labels are not based on this coding system. Nevertheless, given the performance achieved with zero-shot classification on the simplified SewerML labels, mapping to EN13508-2 codes appears feasible either in a zero-shot fashion, by providing the standard information in the prompt context, where context length allows, or by fine-tuning VLMs on standard-compliant data. A further limitation is the small size of the evaluation dataset (100 images), which constrains the statistical robustness of the reported performance metrics. The reported scores should therefore be interpreted as indicative rather than inferential, reflecting exploratory benchmarking rather than

https://doi.org/10.1017/wat.2026.10019 Published online by Cambridge University Press

Table 3. Illustrative examples of model predictions and generated explanations for different sewer defect classes

Image	Class	GPT-4.1 Mini (temp. = 0.0)	o3 (reasoning = low)	Qwen2.5 (32B, (temp. = 0.0))
	RB	**Predicted: RB** […] multiple visible cracks and fractures along the pipe's interior surface […]	**Predicted: RB** […] fissures run irregularly across the pipe […] fractured rather than just being scuffed.	**Predicted: RB** […] significant break or collapse in the pipe structure, with a visible gap […]
	OB	**Predicted: OB** […] surface shows clear signs of wear and erosion, especially on the left side […]	**Predicted: ND** […] only minor roughness and a cleaning nozzle […] do not constitute a defect.	**Predicted: RB** The image shows a significant crack or break in the pipe, […]
	PF	**Predicted: DE** The pipe shows a clear deformation with a pronounced bending or sagging shape, […]	**Predicted: DE** […] it has folded inward creating a pronounced spiral/oval shape that narrows the flow area.	**Predicted: ND** The pipe interior appears smooth and unbroken, with no visible signs of root, […] deformation.
	RO	**Predicted: ND** The pipe interior appears smooth […] with no visible cracks, roots, […]	**Predicted: RO** Fine, hair-like strands protrude from the pipe wall near the 12 o'clock position […]	**Predicted: OB** The pipe surface shows signs of wear and erosion, with visible scratches […]

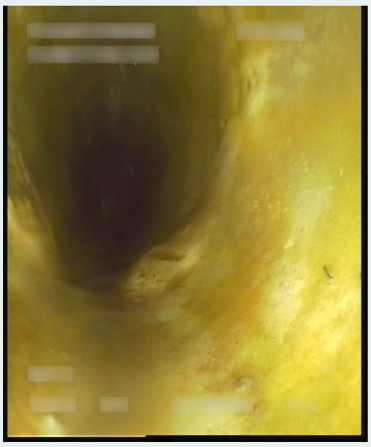
(*Continued*)

definitive model ranking. The SewerML subset we used was not sampled to reflect the natural frequency distribution of classes in real-world inspections. Consequently, performance results should not be directly compared with larger, representative datasets. However, this asymmetry is crucial when designing specialist training datasets. Based on our results, two strategies could be pursued: (a) targeted datasets that compensate for low-performing classes, building on our qualitative assessment, or (b) robustly sampled datasets that reflect field distributions, as recommended by Meijer et al. (2019).

The models were instructed to return a default *ND* class when images were too unclear for classification. This occurred 48 times

https://doi.org/10.1017/wat.2026.10019 Published online by Cambridge University Press

Table 3. (*Continued*)

Image	Class	GPT-4.1 Mini (temp. = 0.0)	o3 (reasoning = low)	Qwen2.5 (32B, (temp. = 0.0))
	DE	**Predicted: DE** The pipe shows a noticeable inward deformation or sagging on the right side […]	**Predicted: ND** Pipe wall appears […] with no visible cracks, roots, or misshapen areas	**Predicted: RB** […] the dark area and the yellowish pipe material suggests a structural defect.

(0.91% of the dataset), mostly from Moondream 1.8B (40 times), which often returned this output for clear images, suggesting hallucination. Llama 4 and 3.2 also returned the default option eight times, but only for genuinely dark or ambiguous images. Given the rarity of default outputs, their influence on overall accuracy is negligible except for Moondream 1.8B. A further limitation lies in the subjectivity of the qualitative assessment. Although definitions were established carefully, their inherently fuzzy nature introduces variability. Future work should involve a committee of trained sewer inspectors to validate textual interpretations and better quantify the added value of descriptive reasoning.

The textual outputs of VLMs provide valuable interpretability, offering insight into model shortcomings and potential improvements. However, this analysis is time-consuming and partially offsets the efficiency gains of automated detection. A practical solution would be to design VLM outputs that combine textual explanations with real-time image highlighting of the perceived defect, e.g., by providing bounding boxes (Situ et al., 2023) or segmentation masks (Zhou et al., 2022). Such a system, if computationally efficient and deployable locally, could substantially reduce the manual inspection load for largely non-defective pipes while supporting rapid defect screening.

Benchmarking a moving target

During the submission and review phase of this manuscript, several new proprietary VLMs were released, including GPT-5, GPT-5.1 and GPT-5.2. Since the experiments reported above were completed prior to their release, we did not repeat the full benchmark across all figures and analyses. Instead, to assess whether our main conclusions remain valid, we reran the same experiment on these newer models with a high reasoning setting, which is expected to provide their strongest performance.

Contrary to expectations based on commonly reported multimodal benchmarks, none of the GPT-5 variants outperformed the earlier reasoning model (o3) or the best non-reasoning model (GPT-4.1 mini) on this task. Specifically, GPT-5 achieved a macro-F1 score of 0.475 (macro recall 0.483, macro precision 0.477), GPT-5.1 achieved a macro-F1 of 0.442 (recall 0.447, precision 0.464) and GPT-5.2 achieved a macro-F1 of 0.449 (recall 0.470, precision 0.513).

These results are noteworthy given that newer model generations typically demonstrate clear gains also on general-purpose vision benchmarks such as Charxiv (Wang et al., 2024) and Screenspot-pro (Li et al., 2025). Our findings indicate that such improvements do not necessarily translate to specialised applications such as sewer defect classification. This reinforces one of the central messages of this study: performance reported on generic vision-language benchmarks is not a reliable proxy for task-specific effectiveness in applied engineering contexts.

More broadly, this highlights the difficulty of defining a stable "state of the art" when foundation models evolve on timescales shorter than the academic review process. Rather than chasing individual model releases, our results argue for systematic, task-driven evaluation and for the development of domain-adapted VLMs whose performance, interpretability and calibration can be validated against application-specific requirements.

Conclusions

In this study, we present the first comprehensive assessment of VLMs for automating the classification of sewer inspection data. We evaluated 18 VLMs on a balanced dataset in a zero-shot setting (i.e., without task-specific training) for multi-label classification of six sewer defect classes. We quantified per-model performance and qualitatively assessed their textual outputs. From this evaluation, we conclude:

- The best-performing VLM, GPT-4.1 mini, achieved an overall macro-F1 score of 0.50. This is a notable finding since OpenAI claims that the much larger o3 model (ranked 2nd in our analysis) represents their most capable vision systems.
- Small- and mid-sized open-source models (e.g., Qwen2.5-VL and Llama 4) performed well above random guessing, with F1 scores > 0.3, indicating clear potential for further improvement through fine-tuning on specialized datasets.
- Model size affected open-source VLMs: performance rose sharply from very small to small models, then increased roughly linearly along a Pareto front. In contrast, OpenAI models showed no size-performance trend, since one of the smallest

https://doi.org/10.1017/wat.2026.10019 Published online by Cambridge University Press

variants performed best. Undisclosed details of OpenAI models preclude overall saturation assessment versus model size.
- Temperature and reasoning levels showed no consistent influence on results, although a more thorough analysis with multiple trials is needed.
- All models performed worst on the Deformation and Production Error classes. Self-reported confidence scores were generally unreliable, showing little difference between correct and incorrect predictions, with the exception of o3 for which higher reasoning effort appeared to improve consistency.
- Analysis of textual explanatory outputs revealed varying levels of hallucination across models and highlighted common sources of misclassification. Better-performing models tended to exhibit fewer complete hallucinations.

Overall, our study highlights pathways for improved automated defect detection while acknowledging the conservative and risk-averse nature of the water sector. Small- and medium-sized open-source models, such as Qwen2.5-VL (32B), are well suited for self-hosted or on-site deployment, allowing sensitive infrastructure data to remain within utility-controlled environments. This setup aligns with security and legal constraints and provides a practical basis for gradual adoption, including further fine-tuning on curated, utility-owned datasets. Future work should also test whether including sewer asset metadata (e.g., material and conduit shape) in the prompt can reduce common misinterpretations, such as classifying egg-shaped conduits as deformed or biofilm on plastic pipes as surface damage. This context is likely to improve performance for the challenging classes of Deformation and Production Error. Another direction lies in coupling VLMs with image segmentation and defect localisation methods, extending outputs to bounding boxes or segmentation masks, as in Google's Gemini 2.5 (Comanici et al., 2025). Such integration would enable models not only to classify and explain defects but also to visually highlight them in real time, further supporting operators. These directions could be combined to deliver a new generation of AI tools for sewer inspection, potentially enabling fully autonomous inspections powered by agentic AI, whether offline, online, or at the edge on inspection robots. However, this would likely require substantially smaller yet highly capable models – potentially below a few billion parameters. Recent studies suggest that approaches operating directly on video streams, rather than single images, may be more suitable for this task (Pai et al., 2025), while model size could be further reduced by adopting highly targeted, task-specific architectures trained using self-supervised learning rather than general-purpose VLMs (Yildizli et al., 2025).

Open peer review. To view the open peer review materials for this article, please visit http://doi.org/10.1017/wat.2026.10019.

Data availability statement. All results, images and labels are hosted on Hugging Face (https://huggingface.co/spaces/rtaormina/vlms_for_sewerml) with an interactive explorer. This also includes results and predictions for the GPT5.x model variants.

Acknowledgements. The authors thank Jip Steiger, Mike van der Voorden, Akrivi Alexandraki and Antía García Rivero for selecting the SewerML subset used in this work. Generative AI tools (i.e., ChatGPT using GPT4o and o3 models) were used to refine the language of the manuscript; all content was verified and remains the sole responsibility of the authors.

Author contribution. Conceptualisation, R.T. and J.A.v.d.W.; methodology, R.T. and J.A.v.d.W.; formal analysis, R.T. and J.A.v.d.W.; software, R.T.; data curation, R.T. and J.A.v.d.W.; writing – original draft, R.T. and J.A.v.d.W.; writing – review and editing, R.T. and J.A.v.d.W.; visualisation, R.T. and J.A.v.d.W.

Funding statement. This research received no external funding. Open access funding provided by Delft University of Technology.

Competing interests. The authors declare no conflict of interest.

References

Bai S, Chen K, Liu X, Wang J, Ge W, Song S, Dang K, Wang P, Wang S, Tang J, Zhong H, Zhu Y, Yang M, Li Z, Wan J, Wang P, Ding W, Fu Z, Xu Y, Ye J, Zhang X, Xie T, Cheng Z, Zhang H, Yang Z, Xu H and Lin J (2025) Qwen2.5-vl technical report. *arXiv Preprint*, arXiv:2502.13923.

CEN (2003) *EN 13508–2 Conditions of Drain and Sewer Systems Outside Buildings–Part 2: Visual Inspection Coding System*. Brussels: European Committee for Standardization.

Comanici G, Bieber E, Schaekermann M, Pasupat I, Sachdeva N, Dhillon I, Blistein M, Ram O, Zhang D and Rosen E (2025) Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv Preprint*, arXiv:2507.06261.

Dirksen J, Clemens FHLR, Korving H, Cherqui F, Le Gauffre P, Ertl T, Plihal H, Müller K and Snaterse CTM (2013) The consistency of visual sewer inspection data. *Structure and Infrastructure Engineering* **9**(3), 214–228.

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N (2020) An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv Preprint*, arXiv:2010.11929.

George A, Shepherd W, Tait S, Mihaylova L and Anderson SR (2025) Explainable deep anomaly detection with sequential hypothesis testing for robotic sewer inspection. *arXiv Preprint*, arXiv:2507.22546.

Granite Vision Team, Karlinsky L, Arbelle A, Daniels A, Nassar A, Alfassi A, Wu B, Schwartz E, Joshi D and Kondic J (2025) Granite vision: A lightweight, open-source multimodal model for enterprise intelligence. *arXiv Preprint*, arXiv:2502.09927.

Haurum JB and Moeslund TB (2020) A survey on image-based automation of cctv and sset sewer inspections. *Automation in Construction* **111**, 103061.

Haurum JB and Moeslund TB (2021) Sewer-ml: A multi-label sewer defect classification dataset and benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13456–13467.

Haurum JB, Madadi M, Escalera S and Moeslund TB (2022a) Multi-scale hybrid vision transformer and sinkhorn tokenizer for sewer defect classification. *Automation in Construction* **144**, 104614.

Haurum JB, Madadi M, Escalera S and Moeslund TB (2022b) Multi-task classification of sewer pipe defects and properties using a cross-task graph neural network decoder. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2806–2817.

Hurst A, Lerer A, Goucher AP, Perelman A, Ramesh A, Clark A, Ostrow AJ, Welihinda A, Hayes A and Radford A (2024) Gpt-4o system card. *arXiv Preprint*, arXiv:2410.21276.

Li K, Meng Z, Lin H, Luo Z, Tian Y, Ma J, Huang Z and Chua T-S (2025) Screenspot-pro: Gui grounding for professional high-resolution computer use. In *Proceedings of the 33rd ACM international conference on multimedia*, 8778–8786.

Meijer D, Scholten L, Clemens F and Knobbe A (2019) A defect classification methodology for sewer image sets with convolutional neural networks. *Automation in Construction* **104**, 281–298.

Meta AI (2024) Llama 3.2: Revolutionizing edge Ai and vision with open, customizable models. Meta Platforms, Inc. Blog post, 25 September 2024. Available at https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/ (accessed 28 July 2025).

Meta AI (2025) The llama 4 herd: The beginning of a new era of natively multimodal Ai innovation. Meta Platforms, Inc. Blog post, 5 April 2025. Available at https://ai.meta.com/blog/llama-4-multimodalintelligence/ (accessed 28 July 2025).

Mistral AI (2025) Mistral small 3.1. Mistral AI. Blog post, 17 March 2025. Available at https://mistral.ai/news/mistral-small-3-1 (accessed 28 July 2025).

Moondream AI (2024) Announcing moondream 2: a 1.8b-parameter mobile-first multimodal model. Moondream AI. Blog post, 22 November 2024. Available at https://moondream.ai/blog/moondream-2 (accessed 28 July 2025).

OpenAI (2024a) Gpt-4.1 overview. Available at https://openai.com/index/gpt-4-1/ (accessed 28 July 2025).

OpenAI (2024b) O3 and o4-mini system card. Technical report. OpenAI. Available at https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf (accessed 28 July 2025).

Opitz J and Burst S (2019) Macro f1 and macro f1. *arXiv preprint*, arXiv:1911.03347.

Pai J, **Achenbach L**, **Montesinos V**, **Forrai B**, **Mees O and Nava E** (2025) *Mimic-video: Video-action models for generalizable robot control beyond Vlas.* arXiv: 2512.15692 [cs.RO]. Available at https://arxiv.org/abs/2512.15692 (accessed 16 March 2026).

Radford A, **Kim JW**, **Hallacy C**, **Ramesh A**, **Goh G**, **Agarwal S**, **Sastry G**, **Askell A**, **Mishkin P and Clark J** (2021) Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PmLR.

Situ Z, **Teng S**, **Feng W**, **Zhong Q**, **Chen G**, **Su J and Zhou Q** (2023) A transfer learning-based yolo network for sewer defect detection in comparison to classic object detection methods. *Developments in the Built Environment* **15**, 100191.

Taormina R and van der Werf JA (2024) Interpretable sewer defect detection with large multimodal models. *Engineering Proceedings* **69**(1), 158.

Team Gemma, Kamath A, **Ferret J**, **Pathak S**, **Vieillard N**, **Merhej R**, **Perrin S**, **Matejovicova T**, **Ramé A and Rivière M** (2025) Gemma 3 technical report. *arXiv Preprint*, arXiv:2503.19786.

Tscheikner-Gratl F, **Caradot N**, **Cherqui F**, **Leitão JP**, **Ahmadi M**, **Langeveld JG**, **Le Gat Y**, **Scholten L**, **Roghani B**, **Rodríguez JP**, **Lepot M**, **Stegeman B**, **Heinrichsen A**, **Kropp I**, **Kerres K**, **Almeida M**, **Bach PM**, **de Vitry M**, **Sá Marques A**, **Eduardo Simões N**, **Rouault P**, **Hernandez N**, **Torres A**, **Werey C**, **Rulleau B and Clemens F** (2019) Sewer asset management–state of the art and research needs. *Urban Water Journal* **16**(9), 662–675.

Wang Y, **Xu X**, **Yu W**, **Xu R**, **Cao Z and Shen HT** (2022) Comprehensive framework of early and late fusion for image–sentence retrieval. *IEEE Multimedia* **29**(3), 38–47.

Wang Z, **Xia M**, **He L**, **Chen H**, **Liu Y**, **Zhu R**, **Liang K**, **Wu X**, **Liu H**, **Malladi S**, **Chevalier A**, **Arora A, and Chen D** (2024) Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems* **37**, 113569–113697.

Wei J, **Wang X**, **Schuurmans D**, **Bosma M**, **Xia F**, **Chi E**, **Le QV, and Zhou D** (2022) Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* **35**, 24824–24837.

Xie Q, **Li D**, **Xu J**, **Yu Z and Wang J** (2019) Automatic detection and classification of sewer defects via hierarchical deep learning. *IEEE Transactions on Automation Science and Engineering* **16**(4), 1836–1847.

Yildizli T, **Jia T**, **Langeveld J and Taormina R** (2025) Self-supervised learning for multi-label sewer defect classification. *Automation in Construction*, **182**, 106751.

Zhai X, **Mustafa B**, **Kolesnikov A and Beyer L** (2023) Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11975–11986.

Zhang J, **Huang J**, **Jin S and Lu S** (2024) Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **46**(8), 5625–5644.

Zhou Q, **Situ Z**, **Teng S**, **Liu H**, **Chen W and Chen G** (2022) Automatic sewer defect detection and severity quantification based on pixel-level semantic segmentation. *Tunnelling and Underground Space Technology* **123**, 104403.

https://doi.org/10.1017/wat.2026.10019 Published online by Cambridge University Press