GENTELL-GAZE: A DATASET AND BENCHMARK FOR LOCALISING PHYSICAL ANOMALIES IN AI-GENERATED VIDEO WITH HUMAN GAZE

Anonymous authors
Paper under double-blind review

ABSTRACT

Generated AI videos are becoming increasingly realistic, making the identification of when and how they diverge from real-world behaviour not only harder, but also more important. An AI video failure that is still relatively common is incorrectly representing physical laws and dynamics, leading to anomalous artefacts in the videos. Correctly identifying and locating these can help create explainable and more trustworthy fake-video detectors, as well as help evaluate vision-language models' ability to understand physics. We introduce the *GenTell-Gaze dataset*: 380 generated videos spanning 80 physics-based scenes and six video generators, with 1,070 manually annotated spatio-temporal anomaly 'tubes', alongside eye-tracking data from 129 participants. We develop *GenTell-Bench*, a benchmark for evaluating vision-language model (VLM) reasoning capabilities. This includes classifying videos as real or AI-generated, as well as recognising, classifying, and localising any present physical anomalies. Using the eye-tracking data, we are able to evaluate whether human visual attention can improve VLM video reasoning and anomaly localisation. Across all models tested, performance falls substantially as the tasks progress from video classification to temporal and spatio-temporal localisation, revealing a persistent gap between recognising generated content and identifying the responsible anomalies with physical reasoning.

1 INTRODUCTION

Increasingly realistic AI-generated videos are becoming commonplace, eroding people's trust in the medium itself. While some fake video detectors reach high detection accuracies, this often comes from *shortcuts*, such as encoding and metadata (Onur Cakiroglu et al., 2026). Such signals are not always available and make the detector's performance generator-dependent (Ni et al., 2026). Physical anomalies, on the other hand, are a signal that is both generator-independent and human-verifiable.

In this work, we focus on the relationship between physical understanding and AI-generated video detection. We ask: **Can models identify visible physical anomalies and use them to correctly detect fake videos?** This is motivated by two reasons. First, it supports *explainable detection* – a model that is able to understand what is suspicious in the video can accompany its verdict with contestable and human-interpretable evidence. Second, it is a natural way to *evaluate physical understanding*, something that is becoming increasingly important as models are incorporated into physical and even safety-critical tasks. Our contributions are as follows:

1. **GenTell-Gaze**, a dataset connecting physical anomalies, human attention, and human judgements on real and AI-generated videos (Figure 1). The combination of these elements enables investigating where failures occur, where people look, and how realistic those videos appear.
2. **GenTell-Bench**, a benchmark to evaluate VLM ability to detect AI-generated videos alongside detecting, classifying, and localising physical anomalies in these videos. It has progressive task difficulty and exposes capabilities that cannot be observed from classification accuracy alone.
3. An indicative picture of current VLM limitations. From the tested models, only GPT-6 Astra achieves better than chance detection accuracy, yet still falls behind human performance. Additionally, we find weak anomaly detection and localisation ability, and substantial sensitivity to prompting and gaze presentation, which shifts models' bias more than their capabilities.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

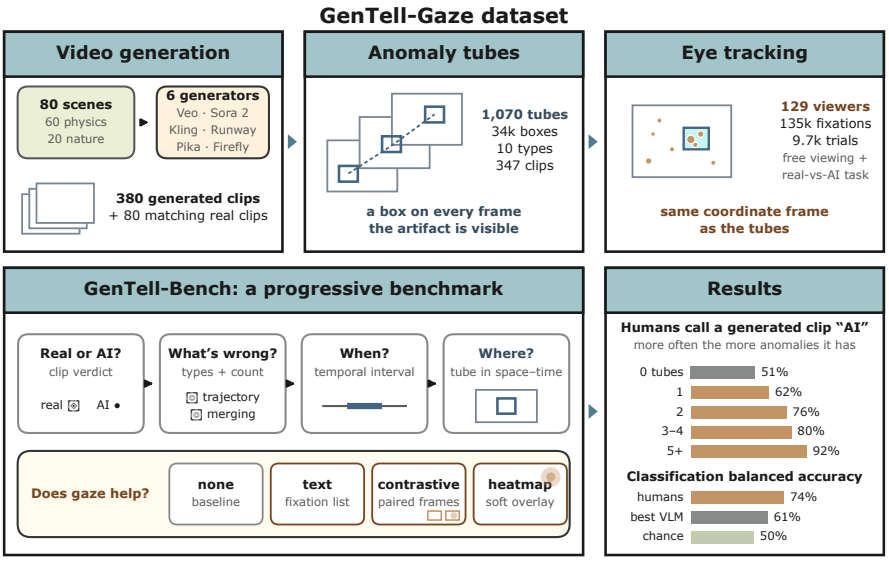


Figure 1: Overview of the GenTell-Gaze dataset and the progressive GenTell-bench benchmark.

2 Background and Related Work

Detecting AI-generated video. Very large corpora of real and AI-generated videos, such as GenVideo (Chen et al., 2026) and GenVidBench (Ni et al., 2026), can now be used to train and test classification accuracy over general scenes. However, detectors trained this way degrade significantly on generators outside their training data (Ni et al., 2026). This suggests that their detection-ability comes mostly from "shortcuts"; indeed, much of a detector's reported accuracy comes from utilising encoding and metadata cues (Onur Cakiroglu et al., 2026). Detection based on physical reasoning may make it more robust, and it has the additional advantage of being human-verifiable. This motivates grounding detection in the physical anomalies that generators produce, which are visible to people and independent of the generator. Skyra (Li et al., 2026b) and DeepTraceReward (Fu et al., 2025) aim to ground their detectors on the presence of physical anomalies, and train their models to output a bounding box and an explanation alongside the real/AI-generated verdict. However, these works did not measure whether the verdict actually followed from the provided evidence, or any other physical-reasoning.

Localising anomalies in generated video. To be able to train and test the connection between detection and physical understanding, we need good ground-truth annotations of physical anomalies within AI-generated videos. Indeed, the annotations of generation errors have grown more detailed, from temporal segments only (Chinchure et al., 2025), to a static box over an interval (Fu et al., 2025; Li et al., 2026b) and, most recently, frame-by-frame annotated regions (Wang et al., 2026; Lin et al., 2025). Results from these works show an intriguing finding: VLMs can recognise generated content, but they are not good at explaining it. In Fu et al. (2025), the models were found lacking at predicting the bounding box location (near-zero box IoU). Tang et al. (2026) finds a similar result, as models tested identified fewer than 10% of the annotated artefacts. Existing work shows that being able to correctly locate anomalies is a task with much room for improvement.

There are several benchmarks and evaluation frameworks aimed at assessing the physical plausibility of generated videos, including (Motamed et al., 2026; Rädsch et al., 2026; He et al., 2025; Cao et al., 2026). Assessing physical plausibility is related to, but distinct from, localising individual visible anomalies that a viewer might notice and use in their judgment. Related video-understanding benchmarks highlight that correctly answering questions about videos does not always indicate strong grounding. For example, Cheng et al. (2026) find that models struggle to identify *when* and *where* the supporting evidence occurs, and Li et al. (2026a) reveal weaknesses in temporal understanding.

Human gaze and synthetic media. It is clear that locating and identifying anomalies in AI-generated videos is still a task at which humans are better than AI. Given that, we might ask if there

is any relevant human-originating signal that we can utilise to improve performance. One such judgment-free signal is eye-gaze. Eye-gaze tracking enables us to record where and when a person's attention was when watching a video. A few works sit at the intersection of eye-gaze tracking and generated AI videos. Gupta et al. (2020) tracked viewers' gaze when watching face-swapped videos to understand which areas help people identify the tampered videos. They found distinct viewing patterns between real and manipulated videos. Shi et al. (2026) also found statistical differences when watching real and AI-generated videos, although these were influenced by the viewer's perception of whether the video was real or fake. Several studies have provided eye-gaze data as input to VLMs, through a trained module (Yan et al., 2024) or directly as markers or heatmaps (Shtedritski et al., 2023; Cai et al., 2024; Yu et al., 2024; Peng et al., 2026). In all of these studies, the eye-gaze data was used to clarify the question rather than as attention cues. Collectively, they found small zero-shot gains, with larger models benefiting more, and heat maps being more helpful than other inputs. Training on how to interpret the gaze data was found helpful (Cai et al., 2024; Yan et al., 2024), but the evidence is far from conclusive. In contrast, StreamGaze, a benchmark to evaluate how effectively models utilize gaze reasoning in streaming videos, did not find that providing gaze input consistently improved reasoning performance (Lee et al., 2025).

3 THE GENTELL-GAZE DATASET

Our dataset consists of several data types. First, it contains a corpus of AI-generated videos, and links to the corresponding real videos. For the AI-generated videos, we provide human-created anomaly annotations, with a spatio-temporal *tube* for each anomaly, as well as a classification of the anomaly type. For all videos, AI-generated and real, we have human-classification accuracy for the video, i.e., the ratio of people who correctly identified it as fake, from an in-person experiment, an online experiment, or both. All videos also have eye-gaze data from 6–29 people from in-person experiments, showing where viewers' attention focused while watching the video. Table 1 presents an illustrative comparison of our GenTell-Gaze dataset compared to other generative-AI detection, anomaly detection and localisation datasets and benchmarks. The organisational structure of the dataset is shown in Figure A.1. We detail each part below.

Table 1: **Datasets and benchmarks comparison.**

Dataset	Matched scenes	Real	AI	Gen.	Anomalies (per video)	Temp.	Spatial	Tube	Types	Human accuracy	Gaze (n)
GenVideo (2024) general scenes; T2V, I2V + web	no	1.2M	1.1M	+19	–	–	–	–	–	–	–
GenVidBench (2025) general scenes; T2V, I2V	prompt/image	0.4M	6.3M	11	–	–	–	–	–	–	–
FakeET (2020) faces; face swap	source clip	331	480	n.r.[a]	–	–	–	–	–	n.r.[d]	40
DeepTraceReward (2025) general scenes; T2V	no	3,318	3,318	7	4,334 (1.3)	✓	static box	–	9	–	–
BrokenVideos (2025) general scenes; T2V	no real	0	3,254[b]	12	n.r.[g]	✓	pixel mask	(∼)	11	–	–
Spotlight (2025) general scenes; T2V	no real	0	600	3	1,604 (2.7)	✓	–	–	6	n/a[h]	–
ViF-CoT-4K / Skyra (2025) general scenes; T2V, I2V	caption / first frame	∼4K paired[c]		10+	n.r.	✓	static box	–	23	–	–
CaC (2026) general scenes; T2V, I2V	no real	0	30,748[b]	+3[e]	n.r.	✓	per-frame box	(∼)	5	–	–
Artifact-Bench (2026) general scenes; T2V	caption	1,350 mixed[c]		+12	–	–	–	–	30	93.6% (4)	–
GenTell-Gaze (ours) physical scenes; I2V	**caption + keyframe**	**80**	**380**	**6**	**1,070 (2.8)**	✓	**per-frame box**	✓	**10**	**74.5% (129)**	**129**

Matched scenes: how generated videos relate to the real ones. *Real/AI*: number of videos; *Gen.*: number of generators; *Anomalies*: human-annotated anomaly instances, with mean per video in brackets; *Temp.*, *Spatial*, *Tube*: whether an anomaly carries a time interval, a spatial region, and a frame-by-frame annotation; *Types*: categories in anomaly taxonomy; *Human accuracy*: real-vs-AI human classification accuracy, with the number of participants in brackets; *Gaze*: number of viewers whose eye movements were recorded. n.r.: not reported. [a] Face-swap manipulations rather than a video generator. [b] All videos are generated: some of the clips are anomaly-free generated videos, which the paper treats as "normal". [c] Real/AI split not reported. [d] Verdicts collected, but accuracy is not reported. [e] Named generators; the paper says "state-of-the-art models" without a count. [g] Anomalies are given as per-frame masks, with 336K frame-level annotations, but without an instance count. [h] Noticeability rated 1–5 rather than real-vs-AI accuracy. (∼): tracked masks; sampling rate n.r. **Comparison is for illustrative purposes only.**

3.1 VIDEO CORPUS

The video corpus contains 380 unique AI-generated videos and 80 corresponding real videos. 360 of the generated videos are derived from 60 Physics-IQ scenes (Motamed et al., 2026) and 20 from 20 nature scenes from Adobe Stock. Real videos were manually selected to include motion and no human faces. This was to test the presence of anomalies when generating scenes with a variety of physical laws, including thermodynamics, fluid mechanics, gravity, collisions, and magnetism.

Video Generation. We used six image-to-video (I2V) generators to create the corpus, all through the Adobe Firefly platform: Google Veo 3.1, OpenAI Sora 2, Runway Gen-4.5, Kling 2.5 from Kuaishou, Pika 2.2, and Adobe's own Firefly video model. The 60 Physics-IQ videos containing distinct scenes were used as sources to prompt AI-video generation by all six generators. The prompt included the scene description text and either the first frame or a key frame of the real video as the visual anchor. The provided image was to ensure that the visual style of the video was aligned with the source clips. Without it, the video's lack of realism may be strongly skewed by artificial-looking rendering rather than misrepresentation of physical laws. This process provided 420 videos, including the real clips. The Adobe Stock videos, containing a mix of wildlife, sports, and other movement, were used to generate 20 more videos, using only Veo as a generator.

We have included multiple generators, as it was shown that the performance of AI-generated video detectors degrades significantly when they are used on videos generated by a model outside of their training data (Ni et al., 2026). Full generation details are in section C.

Video Processing. To prevent a clear distinction between real and AI-generated videos, we cropped every video to 960×720 px and H.264-encoded. Any visible generation watermark was also cropped away. The generated videos derived from Physics-IQ were trimmed to match their real counterpart's duration, 5 s, while Adobe Stock-derived videos were left untrimmed and run longer, with a median of 7.8 s. The file metadata, such as frame rate and file size, was left untouched and can be used as a control in the benchmark.

Video Detection Difficulty. For all videos, AI-generated and real, we provide a human-detection accuracy based on aggregating human judgments from in-person experiments. Each accuracy is calculated based on the judgment of at least 6 distinct viewers, with a median of 18 for real videos, 14.5 for videos generated from Veo, and 8 for videos from other generators. Due to this variation, the accuracies are not directly comparable, but they are a good indication of the video's classification difficulty for a human viewer. The distribution of video difficulty for each generator within the dataset is shown in Figure 2 (a). For 120 videos, including a subset of real, Pika, and Kling videos, we also have detection accuracy from an online experiment, with a higher number of viewers, 59 or more, for each video. Comparing the two detection accuracies, we find that per-clip accuracy in the two settings is moderately correlated ($\rho = 0.60, n = 120$) with no systematic shift (mean difference $+0.005$, 95% CI $[-0.03, +0.04]$). The differences are partly due to sample variability; however, there seems to also be a modest setting-specific component. For more details, see Section D.

3.2 ANOMALY ANNOTATIONS

The next layer in this dataset contains ground truth annotations for where and when in the videos human-perceivable anomalous artefacts are. These were created manually by two annotators on their respective subset of the videos and were then verified by the other. The annotation process was done within a dedicated tool we built for this process. When the annotator identified an anomaly, they created a rectangular bounding box which tracks the artefact over the frames in which it appears. We refer to spatio-temporal bounding boxes as *tubes* for conciseness. The annotator also classifies each of the anomalies into a type, following an ontology adapted from Fu et al. (2025). See Section E for all 10 categories and their description. We add the category *Impossible Physics* to the ontology as a catch-all for anomalies that violate physical laws but do not match any other category.

Overall, the data contains 1,070 annotated anomalies across 347 videos, as 33 AI-generated videos did not contain human-detectable anomalies, with a median of 3 unique anomalies for annotated videos. Figure 2 shows the distribution of the number of anomaly tubes per clip across the dataset (b) and by generator (c). For anomaly types, Object Trajectory is the most common (35%), followed by Object Distortion (20%). Examples of Object Trajectory and Object Distortion are shown in Figure A.4. In terms of spatio-temporal volume, the median anomaly duration is 0.67 seconds, and

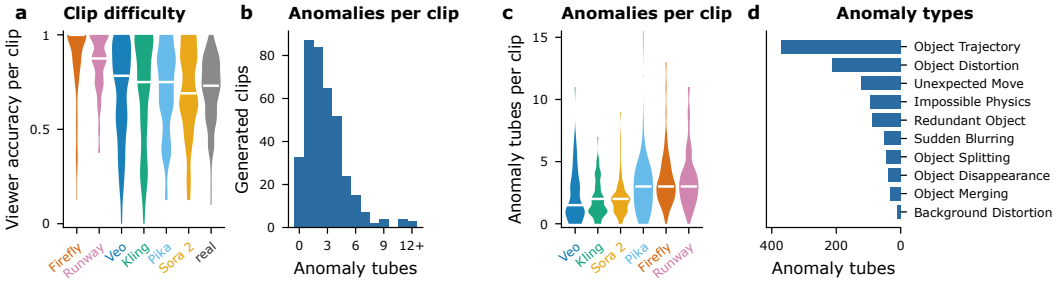


Figure 2: (a) Human detection accuracy per clip from in-person experiments, by generator. (b) Number of annotated anomaly tubes per generated clip; (c) shows the same per generator. (d) Number of anomalies by type across all 1,070 tubes.

the median annotated area is 2.2% of the frame. Per-tube duration, box size and onset time are shown, for each generator, in Figure A.3.

3.3 Eye-gaze Data

Each video, AI-generated and real, has eye-gaze data from at least 6 distinct viewers. All eye-gaze data were collected via facilitated in-person lab experiments at the University of [Anonymised], using a display-based eye-tracker, Gazepoint GP3 (60 Hz). Stimuli were presented on a monitor with a resolution of 1680×1050 pixels, and participants were seated ≈ 60 cm from the monitor. The Gazepoint Control software was used for calibration and the Gazepoint Analysis software to record the gaze data. The collected data comes from three experimental protocols: a free viewing task, in which participants were asked to watch the video and then describe it in one sentence (T1); a judgments task, where participants were asked to watch the video and decide if the view was real or AI-generated (T2); and a two-pass judgments task, where participants watched a video twice before making a judgment (T3). In all experiments, participants were told at the onset that some of the videos they will see are AI-generated.

Real videos have eye-tracking data from all three protocols. The 80 videos generated from Veo have gaze data from T1 and T2 only. All other AI-generated videos have gaze data from T3 alone. See Table A.1 for a summary table of the GenTell-Gaze Dataset, including which of the three eye-tracking experiment protocols are included for each of the generators.

4 The GenTell-Bench Benchmark

GenTell-Bench is a progressive benchmark, that starts from video classification through anomaly recognition, temporal localisation, and full spatio-temporal localisation of video artifacts. The data available in the GenTell-Gaze dataset allows to run the benchmark with eye-gaze data, in three different input types, and test whether it improves the models' performance.

Video Classification. The first task in GenTell-Bench is to judge whether a clip is real of AI-generated. The individual-clip answers are then used to calculate the model's accuracy on *Real* videos, and on *AI* videos. As the dataset itself is imbalanced, with 380 AI and 80 real videos, we also report the model's balanced accuracy. This is important, as calculating the accuracy on the full GenTell-Gaze dataset, which has more AI than real clips, will favour models that are biased towards voting AI, even when they do not have a true ability to distinguish between real and AI-generated videos. The balanced accuracy (BA) removes such bias effects. BA is the mean of the accuracy on real and generated videos; hence, only a value greater than 0.5 is better than chance.

Anomaly Detection Ability. The benchmark evaluates a model's ability to correctly identify and understand anomalies in the videos in a progressive manner: anomaly recognition, anomaly-type classification, temporal localisation, and spatio-temporal localisation. The most basic measure of anomaly recognition is anomaly-count. It reports the mean of the per-clip absolute difference between predicted and ground-truth anomaly count, independently of correctly identifying the type of anomaly. Anomaly-type detection measures the model's ability to correctly identify the anomaly

type, which requires recognising which physical violation has occurred. There are 10 anomaly categories, and the primary metric we consider here is category-presence macro F1. For each of the ten anomaly categories, we compute set-level true positives, false positives, and false negatives from category presence in each clip, then macro-average F1 across categories.

Temporal Localisation. Here we test the model's ability to correctly identify *when* during the video the anomaly occurred. Predicted and ground-truth temporal intervals are compared using temporal intersection-over-union (tIoU) with F1 reported at tIoU thresholds of 0.3. We evaluated both untyped and typed matching. Untyped matching ignores anomaly category, while typed matching is performed only on anomalies where the model correctly detected the anomaly type. In the main results, we report the untyped results to separate anomaly localisation ability from agreement with our defined category schema. Model-produced temporal annotations are matched one-to-one to the human annotations using maximum-cardinality matching, recomputed independently for each of the thresholds (0.3 and 0.5). Further details of metrics and their definitions may be found in Section F.6.

Spatio-Temporal Localisation. The strictest level measures the ability to create a spatio-temporal tube that tracks the anomaly across the video. Models may generate bounding boxes at the beginning, end, and optionally intermediate points of an anomaly. These boxes are linearly interpolated through time in post-processing and compared with the frame-level ground-truth annotations. We define *TubeIoU* as temporal IoU, multiplied by the duration-weighted mean spatial bounding-box IoU, over the temporal intersection, requiring agreement in both when and where an anomaly occurs. As with temporal localisation, we report the untyped F1 score at a threshold of 0.3.

All anomaly detection tasks are conducted on both real and AI-generated videos. On real videos, any detection and tube indicated by the model is considered a false positive. All rates are computed from per-clip valid statistics, and summed over the complete cohort. Precision, recall and F1 are micro-averaged, i.e, the true positives, false positives and false negatives are summed over all valid clips in the cohort, including real clips, where every predicted tube is a false positive, and a single number is computed from the totals. This means that each annotated tube carries equal weight, rather than each clip.

4.1 Gaze-Conditioned Variants

GenTell-Bench can be run with or without gaze as an input. With gaze, we provide the models with gaze data while keeping the videos, anomaly ontology, prompts, and evaluation procedure fixed. Input gaze data is taken from participants' first viewing of the T3 experiment, where they were asked to classify each video as real or generated after the second viewing. We evaluate three gaze-input variants in addition to the no-gaze variant. Each variant gives the gaze data to the model via a different input. The **Heatmap** overlays an eye-gaze heatmap directly on each sampled frame. Raw fixation locations are spatially smoothed using a Gaussian kernel, with the blur scale chosen to reflect approximately two degrees of expected eye-tracker error (Yan et al., 2024). The **Contrastive** input builds on this representation by presenting the unmodified frame alongside a second copy containing the eye-gaze heatmap. This preserves an untouched visual input while still providing a spatially aligned representation of human attention, reducing the risk that modifying the frame itself interferes with visual features learned from unannotated images. Finally, the **Text** input represents gaze using a language-based scanpath format, with temporal information included so that gaze observations can be associated with the corresponding video frames (Mondal et al., 2026).

5 Results

We test our benchmark on the 360-clip physics subset from the GenTell-Gaze dataset on 4 open-weight models, GLM-4.6V-Flash (Hong et al., 2025), InternVL3.5-8B (Wang et al., 2025), Qwen3-VL-8B, and Qwen3-VL-32B (Bai et al., 2025) and the current OpenAI frontier model, GPT-6 Astra (OpenAI, 2026). Full implementation details are in Section F.

The main results can be found in Table 2. Within the detection task, results varied significantly in terms of accuracy on AI-generated or real clips. However, this diversity was mostly due to a model's difference in *bias* for voting *Real* versus *Fake*. In terms of classification ability, only Astra was able to show notable improvement over 50% chance with a balanced accuracy of 61%, which is still well below the human baseline of 74%. The remaining open-weight models sit around chance on the

Table 2: **GenTell-Bench results**.

Model	Gaze input	Real	AI	BA	Count MAE ↓	Temporal F1 ↑	Tube F1 ↑
Without gaze input							
Humans (129 viewers)	–	73	76	**74**	–	–	–
GPT-6 Astra	–	100	22	61	2.24	14.2	**7.4**
GLM-4.6V-Flash	–	33	60	47	**2.18**	**16.0**	1.2
InternVL3.5-8B	–	18	86	52	8.42	4.6	0.0
Qwen3-VL-32B	–	71	29	50	4.12	11.5	1.1
Qwen3-VL-8B	–	80	19	49	3.18	4.4	0.8
With gaze input							
GLM-4.6V-Flash	text	37	65	51 (+9%)	2.05 (–6%)	18.3 (+14%)	1.4 (+11%)
	contrastive	8	83	46 (–2%)	2.20 (+1%)	16.4 (+2%)	0.2 (–86%)
	heat map	10	85	48 (+2%)	**2.03** (–7%)	**20.8** (+30%)	**2.2** (+73%)
InternVL3.5-8B	text	27	86	**56** (+8%)	4.54 (–46%)	14.0 (+203%)	0.2 (–)
	contrastive	0	99	49 (–5%)	8.95 (+6%)	3.7 (–20%)	0.0
	heat map	19	82	50 (–3%)	10.30 (+22%)	4.5 (–2%)	0.0
Qwen3-VL-32B	text	57	41	49 (–3%)	4.92 (+19%)	10.3 (–10%)	0.6 (–48%)
	contrastive	53	58	55 (+11%)	3.94 (–4%)	13.1 (+13%)	0.0 (–100%)
	heat map	40	55	48 (–5%)	5.06 (+23%)	14.3 (+24%)	1.0 (–16%)
Qwen3-VL-8B	text	83	13	48 (–3%)	3.41 (+7%)	3.3 (–26%)	0.1 (–81%)
	contrastive	87	12	50 (–)	2.59 (–19%)	2.8 (–36%)	0.0 (–100%)
	heat map	58	40	49 (–)	3.09 (–3%)	7.5 (+69%)	0.7 (–6%)

Real and *AI* are recall on real and generated clips and *BA* their mean (balanced accuracy), in %; *Count MAE* is the mean absolute error in the number of anomalies per clip; *Temporal* and *Tube F1* are localisation F1 (%) at IoU 0.3 against the human-annotated tubes. Bold: best per column; underline: second best, within each block. Parentheses give the change from the same model's no-gaze run, relative in %. Green indicates improvements and red deteriorations. Human row includes 3,346 judgements from in-person experiments.

balanced accuracy. Astra and both Qwen models were biased towards answering *Real*, whilst GLM and Intern were biased more towards classifying a video as AI-generated.

Anomaly detection. Focusing on anomaly detection, the model with the lowest count error was GLM (2.18), with Astra close behind (2.24). The same pattern occurs for temporal detection, with GLM leading (16.0), and Astra close behind (14.2). For Tube localisation, however, Astra was the only model that showed any capability (7.4). The relationship between anomaly detection and classification was incredibly strong. Essentially (99.99%), for all models, under all conditions, *if the model identified 1 or more anomalies, it classified the video as AI-generated* (see Table A.4). However, the reverse is not true. As shown in Figure 3 a, all models, except Qwen-32B, did classify clips as AI videos even when they did not count any anomalies. The models biased towards answering *Real* did this rarely (9/300 for Astra and 5/300 for Qwen-8B). However, Astra was correct in all of these instances, which Qwen-8B was not. If this logic had been matched with a better anomaly-detection ability, these models would have been better detectors. Indeed, in Figure 3 b we see that Astra undercounts anomalies 92% of the time (and never overcounts), leading to its real-leaning bias. The Qwen models do overcount, with 7% for the 8B model and 16% for the 32B one.

GLM was the best in terms of anomaly counting, meaning it was the best at estimating the correct number of anomalies per clip (note this compares only counts, not if each counted anomaly corresponds to a real one), but its detection capability was undermined by classifying videos with no detected anomalies as AI-generated. Intern, the most AI-biased model, was overzealous both in terms of assigning videos with no detected anomalies as AI, and in over-assigning anomalies. It is interesting to look at how model classification compares to the count of the ground-truth anomalies (Figure 3 c). For human viewers, there is a statistically significant moderate correlation ($\rho = 0.42$,$p < 0.001$) between the *Fake* verdict and the number of anomalies. For the models, Astra shows a small to moderate, significant correlation ($\rho = 0.25$, $p < 0.001$); Intern and Qwen-32B show an even smaller one ($\rho = 0.15$, $p = 0.006$, and $p = 0.076$, respectively); and the rest none.

Anomaly Classification. In addition to recognising where and when something went wrong, we can also ask *what* went wrong. We do this with a 10-category annotation-type classification. For

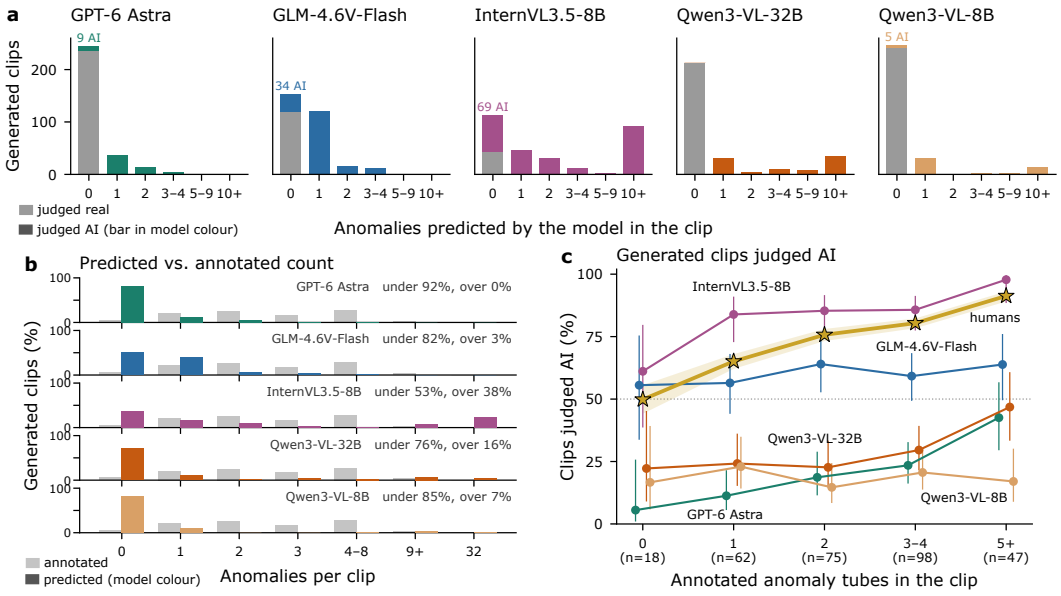


Figure 3: **Verdicts and anomaly counts**. **(a)** Verdict against the model's own predicted anomaly count; any listed anomaly leads to an AI-generated verdict. **(b)** Predicted (colour) versus annotated (grey, fixed) counts per clip, with the share of clips under- and over-counted. **(c)** Share of clips judged AI against the human-annotated count (95% Wilson band). No-gaze runs only.

clip-level category F1, i.e., correctly naming an anomaly type present within the clip, Astra preforms best, with 21.5%, with the best open model, Intern, achieving 12.7%. When judging a clip as AI-generated, Astra names at least one anomaly type correctly in 73% of cases (47 of 64 clips), but recovers the complete set in just 4 clips. However, 85% of the types it does name are present, so most of its errors are due to omissions rather than misidentification. When it correctly annotated a tube, the type is also right in 72% of cases. In comparison, the open models manage to correctly name the type of anomaly only on 22–49% of the clips they judged as fake. The models also differed in how well they were able to recognise different categories: Astra reaches 39–47% F1 on the three discrete identity categories —Object Splitting, Object Merging and Object Disappearance—but only 11% on Object Trajectory, the most common type of anomaly. For the open models, Qwen-32B has the best per-category performance, with 31% on Object Distortion, followed by Intern, with 29% on Unexpected Move. See Figure A.6 for full results.

Reasoning. Astra's output included the model's reported reasoning (not real thinking trace), which helped uncover a very interesting phenomenon. In many cases, Astra noticed an irregularity, and even outright said there was an anomaly, but then "convinced itself" there was a *real-world* justification. For example: "there are odd trajectories, like how the tennis ball reverses twice" → "It seems like it might involve a trick with physics or magnets" (`sora2/0198`, 100% judged AI by humans). Or: "odd shadows resembling rods between frames 18–28" → "It looks like the balls might be on strings" (`pika/0180`, 88% judged AI by humans). Comparing Astra's performance to that of humans (see Figure A.7), we notice two interesting findings. First, Astra's verdicts were aligned with human difficulty ($\rho = 0.34$, $p < 0.001$), whereas the three open models showed only a weak correlation ($\rho \approx 0.12$–0.16). In addition, Astra's average response length *increased* for videos humans found easier, which makes sense when combined with what we noticed about Astra's reasoning.

We can therefore ask the question, if Astra was less conservative in its *Fake* judgments, would it have performed better? The answer is yes, a bit. If we "re-judge" as AI-generated based on the reasoning output, not the verdict, we can get better performance (Table A.6). Adding twelve missed clips where Astra stated outright that a defect was found and yet concluded "real", lifts balanced accuracy to 62.8% with no accuracy drop on real videos. Adding clips with a summary that explicitly states undecided lifts the performance further to 65.8 % balanced accuracy, but at the cost of 5 false positives on real clips.

Adding Gaze as input. We tested providing eye-gaze to the models via three input modes—as a text list, as contrastive frame pairs, and as a heat-map—to all the models. For the open models we tested it on a complete run, and for Astra, a 120-clip, difficulty-distribution-matched subset. As shown in Table 2, no type of gaze-input led to consistent improvement across models. The highest improvements in detection were for Intern, going from 52% to 56% with gaze input as text, and Qwen-32B, which went from 50% to 55%. In both of these cases, there is also noticeable improvement in the anomaly count and temporal localisation. The best results for count, temporal and spatial location, were for GLM with heatmap, all showing improvement. The count and temporal F1, which GLM was already best at prior to gaze input, improved by 7% and 30%, respectfully. The tube F1 improved by 73% to 2.2, but was still far below Astra's baseline result of 7.4. GLM was the only model that showed improvement in all metrics under two gaze condition, however, the improvement with text input was lower. For most models, the gaze input changed the level of bias shift towards voting AI, manifesting both as an increase both in judging more clip without idetified anomalies as *Fake* (Figure A.8) and in an increase in the overestimation of anomalies (Figure A.9). Combined, these over-shadowed any gains in anomaly detection and localisation.

For GPT-6 Astra, the findings were different. The results Table A.8 remained largely unchanged, and even a bit worse compared to the non-gaze results on the same 120-clip subset. Notably, analysing the reasoning shows that in a zero-shot setting, Astra largely ignores the gaze-input. The tubes it reported remained the same, and its reasoning refers to the gaze only in 1–2 % of cases for the textual and contrastive inputs and in 18 % for the heat map. For the latter, the reference was mainly to ignore the evidence, or the model misread the gaze as image content.

Prompt Sensitivity. To test the sensitivity of conducting the benchmark with different prompts, we ran GenTell-Bench on a distributionally similar clip subset (120 clips, same real-to-AI ratio) for the open models with three additional prompts. Results were sensitive to the choice of prompt, but not in manner that changes the results qualitatively. As shown in Section G.6, varying the configuration space (prompt+gaze impact) on the open models can lead to a small improvement in detection-ability, but the main impact is on shift the bias to respond *Fake* or *Real*. The most striking example is Intern, that shifts from being very AI-biased in some configurations to very real-biased in others. A full analysis of the prompt sensitivity can be found in Section G.6.

This sensitivity analysis also helps us better understand where gaze data is helpful, as only one metric shows some consistent, albeit small, improvement. Temporal localisation improves slightly under gaze input. Across the 48 prompt variations x gaze input forms, tested only for the open models, the temporal F1 is higher with gaze than without in 26 (mean $+1.8pp$), and in 23 of the 33 cells once the five combinations that collapse anomaly output are set aside (mean $+2.6pp$).

6 Conclusion

This work investigates AI model ability to detect AI-generated videos through physical anomalies, offering a concrete test of physical understanding. We bring together spatio-temporal anomaly annotations, human judgements and eye-gaze, enabling the evaluation of detection accuracy alongside the ability to identify what went wrong, when and where. Our results show that modern VLMs are currently limited in both detection and anomaly localisation, with the strongest classifier still substantially below human performance. Our analysis offers unique insights into the underlying capabilities, such as which models are strong at identifying which anomaly types. In a zero-shot setup, human gaze provides small gains in temporal localisation, but no consistent improvement across tasks. Our findings highlight there is still much room for improvement in VLM physical understanding, and its relationship to reasoning about the authenticity of videos. We believe the GenTell-Gaze dataset and GenTell-Bench can help improve and test not only how well models can detect generated videos, but how well they can ground their decisions in sound physical understanding that people can inspect and verify.

AI use statement

In this work, we used generative AI tools to implement methods (the benchmark harness, inference clients and evaluation code), to clean and reformat the dataset (the scripts that build and verify the release tables), and to provide feedback on the experimental design. We have not used generative AI

tools to generate synthetic data, develop theoretical models or conceptual frameworks, formulate or prove mathematical claims, propose hypotheses, translate, or support qualitative analysis. We have not used them to produce any annotation or gaze data: all anomaly annotations were drawn manually by human annotators and all gaze data were recorded from human participants. Additionally, we used generative AI tools to create and edit software code, including for generating some of the figures for the paper and appendix. Paper writing was done by the authors without the use of generative AI, however, it was used to help draft some of the text in the appendix, in which case it was reviewed and rewritten by the authors. We have reviewed all AI-assisted work. AI-written code was checked by an automated test suite, and every released table is validated against its sources by a verification script; reported numbers were reproduced from the raw predictions by a single table-building script. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All human-subject data in GenTell-Gaze—eye-tracking recordings and real-vs-AI judgements from the in-person sessions, and judgements from the online study—were collected under a protocol approved by the ethics committee of the authors' institution (details withheld for anonymity and to be provided on acceptance). Every in-person participant gave informed written consent before taking part and could withdraw at any time. In-person participants were compensated for their time with a gift voucher; online participants were entered into a prize draw. No personal or demographic data was collected. participants appear in the dataset only as arbitrary identifiers, and gaze coordinates are expressed relative to the video frame.

The video stimuli were selected to be ethically neutral. None of the 460 clips shows an identifiable person, a face, or a scene with political, violent, sexual or otherwise sensitive content; the source material consists of physical scenes (falling, rolling and colliding objects, fluids, magnets and thermal processes) and motion-rich nature footage. This was a purposeful choice, as it keeps the benchmark focused on physical plausibility rather than on the identity or deepfake artifacts that dominate face-centred datasets.

GenTell-Gaze contains AI-generated video. We release these clips solely for the purpose of detecting and localising generation artifacts, and we do not release the prompts or seeds needed to regenerate them at scale. We recognise the dual-use character of any detection benchmark—knowing which artifacts humans and models notice could in principle guide generators toward hiding them—but the decision to not include human-related content was done to minimise potential misuse. In addition, the dataset set is a good size of a benchmark, but not significant enough for training.

REPRODUCIBILITY STATEMENT

The dataset is larger than the Supplementary Material limit. For review, it can be found in a private link `https://figshare.com/s/ad4d02c48cda71c02acc`. Upon acceptance, we will release the GenTell-Gaze Dataset publicly alongside this paper: the generated videos, frame-level anomaly annotations, and eye-tracking fixations with per-trial judgements. Appendix A documents its file structure, join keys and coordinate and timing conventions. The benchmark protocol is specified in full across the appendix:

- model checkpoints pinned to exact revisions, together with serving, decoding and hardware settings (Appendix F.2);
- frame sampling for input into models (Appendix F.1);
- the exact prompts and the JSON output schema enforced during decoding (Appendix F.5);
- metric definitions, including f1 score, temporal IoU, and tube IoU (Appendix F.6);
- the fixed 120-clip subset used for the prompt-sensitivity and Astra gaze input experiments (Appendix G.6).

Open-weight models were ran with zero temperature and a fixed seed. Small run-to-run differences nonetheless arise from floating-point non-determinism across GPU types and request batching, which we quantify in Appendix G.6. GPT-6 Astra is accessed through a proprietary API that

exposes a single undated snapshot and no temperature control, so its outputs cannot be exactly regenerated; we report its settings in Appendix F. Full implementation code will be shared with the accepted paper.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. Vip-llava: Making large multimodal models understand arbitrary visual prompts. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12914–12923. IEEE, 2024.

Yu Cao, Ziquan Liu, Zhensong Zhang, Jiankang Deng, Shaogang Gong, and Jifei Song. Each judge its own yardstick: Discovering per-vlm taxonomies for physical video evaluation. *arXiv preprint arXiv:2606.22918*, 2026.

Haoxing Chen, Yan Hong, Zizheng Huang, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Jun Lan, Huijia Zhu, Jianfu Zhang, Weiqiang Wang, et al. Demamba: Ai-generated video detection on million-scale genvideo benchmark. *Science China Information Sciences*, 69(6):162103, 2026.

Zixu Cheng, Jian Hu, Ziquan Liu, Chenyang Si, Wei Li, and Shaogang Gong. V-star: Benchmarking video-llms on video spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9155–9164, 2026.

Aditya Chinchure, Sahithya Ravi, Pushkar Shukla, Vered Shwartz, and Leonid Sigal. Spotlight: Identifying and localizing video generation errors using vlms. *arXiv preprint arXiv:2511.18102*, 2025.

Xingyu Fu, Siyi Liu, Yinuo Xu, Pan Lu, Guangqiuse Hu, Tianbo Yang, Taran Anantasagar, Christopher Shen, Yikai Mao, Yuanzhe Liu, et al. Learning human-perceived fakeness in ai-generated videos via multimodal llms. *arXiv preprint arXiv:2509.22646*, 2025.

Parul Gupta, Komal Chugh, Abhinav Dhall, and Ramanathan Subramanian. The eyes know it: Fakeet-an eye-tracking database to understand deepfake perception. In *Proceedings of the 2020 international conference on multimodal interaction*, pp. 519–527, 2020.

Xuan He, Dongfu Jiang, Ping Nie, Minghao Liu, Zhengxuan Jiang, Mingyi Su, Wentao Ma, Junru Lin, Chun Ye, Yi Lu, et al. Videoscore2: Think before you score in generative video evaluation. *arXiv preprint arXiv:2509.22799*, 2025.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Daeun Lee, Subhojyoti Mukherjee, Branislav Kveton, Ryan A Rossi, Viet Dac Lai, Seunghyun Yoon, Trung Bui, Franck Dernoncourt, and Mohit Bansal. Streamgaze: Gaze-guided temporal reasoning and proactive understanding in streaming videos. *arXiv preprint arXiv:2512.01707*, 2025.

Baiqi Li, Kangyi Zhao, Ce Zhang, Chancharik Mitra, Jean de Dieu Nyandwi, and Gedas Bertasius. Timeblind: A spatio-temporal compositionality benchmark for video llms. *arXiv preprint arXiv:2602.00288*, 2026a.

Yifei Li, Wenzhao Zheng, Yanran Zhang, Runze Sun, Yu Zheng, Lei Chen, Jie Zhou, and Jiwen Lu. Skyra: Ai-generated video detection via grounded artifact reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4482–4493, 2026b.

Jiahao Lin, Weixuan Peng, Bojia Zi, Yifeng Gao, Xianbiao Qi, Xingjun Ma, and Yu-Gang Jiang. Brokenvideos: A benchmark dataset for fine-grained artifact localization in ai-generated videos. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13421–13427, 2025.

Sounak Mondal, Dimitris Samaras, Gregory Zelinsky, and Minh Hoai. Gaze-to-text generation: Beyond categorical decoding of human attention. In *European Conference on Computer Vision*, pp. 411–430. Springer, 2026.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 948–958. IEEE, 2026.

Zhenliang Ni, Qiangyu Yan, Mouxiao Huang, Tianning Yuan, Yehui Tang, Hailin Hu, Xinghao Chen, and Yunhe Wang. Genvidbench: A 6-million benchmark for ai-generated video detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 15582–15590, 2026.

Mert Onur Cakiroglu, Zhihe Lu, Mehmet Dalkilic, and Hasan Kurban. Auditing generalization in ai-generated video detection: A six-control protocol and the vidaudit toolkit. *arXiv e-prints*, pp. arXiv–2606, 2026.

OpenAI. GPT-6 Astra: A new generation of intelligence. `https://openai.com/index/gpt-6-astra/`, 2026.

Taiying Peng, Jiacheng Hua, Miao Liu, and Feng Lu. In the eye of mllm: Benchmarking egocentric video intent understanding with gaze-guided prompting. *Advances in Neural Information Processing Systems*, 38, 2026.

Tim Rädsch, Yuki M Asano, Hilde Kuehne, Stefan Bauer, Priyank Jaini, Robert Geirhos, and Carsten T Lüth. Physics-iq verified. *arXiv preprint arXiv:2606.18943*, 2026.

Danqing Shi, Lan Jiang, Katherine M Collins, Shangzhe Wu, Ayush Tewari, and Miri Zilka. Perceived authenticity drives gaze behavior when watching ai-generated videos of physical scenes. *Scientific Reports*, 2026.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11953–11963. IEEE, 2023.

Yuqi Tang, Yang Shi, Zhuoran Zhang, Qixun Wang, Xuehai Bai, Yue Ding, Ruizhe Chen, Bohan Zeng, Xinlong Chen, Xuanyu Zhu, et al. Artifact-bench: Evaluating mllms on detecting and assessing the artifacts of ai-generated videos. *arXiv preprint arXiv:2605.18984*, 2026.

Jiyuan Wang, Huan Ouyang, Jiuzhou Lin, Chunyu Lin, Dewen Fan, Boheng Zhang, Haonan Fan, Fei Zuo, Jia Sun, Huaiqing Wang, et al. Cac: Advancing video reward models via hierarchical spatiotemporal concentrating. *arXiv preprint arXiv:2605.11723*, 2026.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Kun Yan, Zeyu Wang, Lei Ji, Yuntao Wang, Nan Duan, and Shuai Ma. Voila-a: Aligning vision-language models with user's gaze attention. *Advances in neural information processing systems*, 37:1890–1918, 2024.

Runpeng Yu, Weihao Yu, and Xinchao Wang. Attention prompting on image for large vision-language models. In *European Conference on Computer Vision*, pp. 251–268. Springer, 2024.

A Dataset Structure

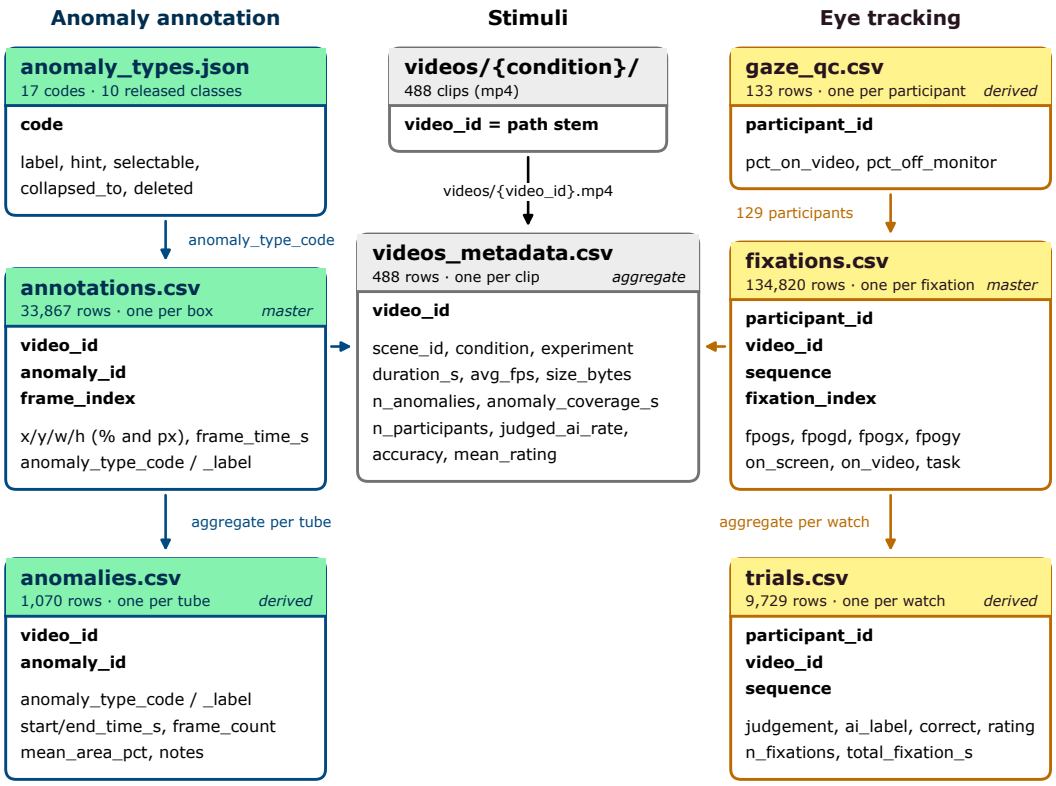


Figure A.1: **Structure of the GenTell-Gaze release.** Every table joins on `video_id`, the clip's path under `videos/` without extension. The annotation track (left) stores one bounding box per annotated frame in `annotations.csv`, aggregated to one row per anomaly tube in `anomalies.csv`; the ontology in `anomaly_types.json` maps the annotator's 17 codes to the 10 released classes. The eye-tracking track (right) stores one row per fixation in `fixations.csv`, in video-frame coordinates, aggregated to one row per participant × clip × watch in `trials.csv` together with the real-vs-AI judgement; `gaze_qc.csv` gives per-participant gaze-loss proportions. `videos_metadata.csv` holds one row per released clip with file properties and per-clip aggregates of both tracks. Bold columns are keys; arrows into `videos_metadata.csv` are `video_id` joins. The two master tables are authoritative and the derived tables are regenerated from them.

The GenTell-Gaze Dataset combines videos, spatio-temporal anomaly annotations and eye-tracking data. The Figure A.1 shows the file structure relating these components. The two master tables, `annotations.csv` and `fixations.csv`, store the bounding boxes per frame and individual fixations, respectively. Derived tables `anomalies.csv` and `trials.csv` summarise each anomaly tube from annotations and single watch from experiments, while `videos_metadata.csv` records properties per-clip and aggregates from both annotation and eye-tracking data. The tables join on `video_id` which is the path stem of each clip within the `videos/` directory, without the file extension. Videos are organised by condition, which identifies the generator or real-video source.

The dataset contains 460 unique clips: 380 generated videos and 80 real source videos. The metadata describes 488 video versions because 28 real scenes have different temporal cuts across experiments. Distinct identifiers allow gaze records to be linked to the temporal cut shown in each experiment. The GenTell-Gaze Benchmark uses the content and temporal versions from experiment three.

At paper acceptance, we will release the 380 generated videos, which we produced ourselves, together with the anomaly annotations and gaze data. Real source videos are not included in this video

release, but their metadata and associated experimental records are retained. The forthcoming public repository will include a single reconstruction script to retrieve and process the 60 physics videos from the Physics-IQ repository and the 20 Adobe Stock videos, reproducing the real source clips and experiment-specific temporal cuts used in the gaze experiments and GenTell-Gaze Benchmark.

The file contains 33,867 bounding boxes, with one row per anomaly and annotated frame. These rows form 1,070 contiguous anomaly tubes in `anomalies.csv`. Each tube has a category, an inclusive start and end frame, and summary geometry. The taxonomy file, `anomaly_types.json`, maps the annotation tool's codes to the ten released classes. Boxes are represented by their top-left position, width and height as fractions of the video frame. Frame indices are zero-based, and timestamps correspond to the decoded frames. This preserves alignment across videos with different frame rates.

Eye movements were recorded with a Gazepoint GP3 tracker at 60 Hz and processed into fixations by Gazepoint Analysis. Videos of 960 × 720 pixels were displayed at 1400 × 1050 pixels on a 1680 × 1050 monitor, leaving a 140-pixel vertical black bar on each side. We transformed the exported screen coordinates into video-normalised coordinates, matching the reference frame of the bounding boxes. The resulting fixations.csv contains 134,820 fixations. The fields fpogx and fpogy give spatial position; fpogs gives the start time in seconds from the onset of the corresponding watch; and fpogd gives fixation duration. The sequence field distinguishes the first and second watches.

Fixations outside the video are retained, with `on_video` identifying those within its displayed bounds. The recording window can also extend beyond playback into the response period. Analyses of attention during playback should therefore select on-video fixations and restrict their temporal overlap to the clip interval. The file `trials.csv` aggregates fixations into one row per participant, clip and watch, preserving experiment and task identifiers. These identifiers distinguish free viewing from classification and single from repeated viewing. Structured judgements and realism ratings are included; full verbal transcripts are outside the release scope. The file `gaze_qc.csv` provides participant-level and aggregate gaze summaries. Online judgements are stored separately as per-clip aggregates in `videos_metadata.csv` and do not appear as rows in the fixation or trial tables.

B Dataset Summary Table

Table A.1: **Summary of GenTell-Gaze dataset Experiments.** This table lists the eye-tracking experiment protocol which was used to generate the corresponding gaze data. *Viewers* is the number of distinct participants with gaze data per clip (min–max); *Judgements* counts real-vs-AI verdicts, and *Acc.* is their corresponding calculated accuracy.

			Anomaly annotation			Eye tracking				
Condition	Exp.	Clips	Tubes	Boxes	Med./clip	Partic.	Viewers/clip	Fixations	Judgem.	Acc.
Veo 3.1	1, 2	80	181	3,719	1.5	80	19–20	27,724	1,586	0.71
Sora 2	3	60	128	5,704	2	49	7–9	13,399	974	0.70
Kling 2.5	3	60	118	3,228	2	49	7–9	13,015	961	0.69
Runway	3	60	221	6,813	3	49	7–9	13,175	967	0.85
Pika 2.2	3	60	205	7,889	3	49	6–9	12,951	952	0.70
Firefly	3	60	217	6,514	3	49	6–9	13,168	965	0.86
Real	1, 2, 3	80	–	–	–	129	18–29	41,388	2,547	0.74
Total	1–3	460	1,070	33,867	–	129	6–29	134,820	8,952	0.74

The Table A.1 summarises the stimulus, annotation and eye-tracking coverage. The dataset contains 460 unique clips, including 380 generated videos, with 1,070 anomaly tubes represented by 33,867 frame-level boxes. Eye-tracking data comprise 134,820 fixations from 129 distinct participants who completed their assigned sessions. The three experiments involved 40, 40 and 49 participants, respectively, with no participant overlap between them.

Experiment one used 40 scenes: 20 from Physics-IQ and 20 from Adobe Stock, each represented by a real clip and a Veo generation. Participants completed two tasks on separate sets of clips. In Task 1, they viewed each clip naturally and described it afterwards, without making a real-versus-AI judgement. In Task 2, they watched once and classified the clip as real or generated. Experiment two used 40 additional Physics-IQ scenes, again with real and Veo versions, and combined Task 2 with Task 3. In Task 3, each clip was viewed twice before a single classification response. Experiment three covered all 60 physics scenes using five generators and real footage, yielding 360 clips, and used Task 3 throughout. Participants were also asked to rate realism on a five-point scale when they judged a clip generated. Gaze was recorded throughout each viewing.

Experiments one and two used four counterbalanced groups, with each participant assigned one version of each scene and 20 clips per task. Experiment three used six fixed playlists, each containing 60 clips: one per scene and ten per condition. Presentation order was fixed within each group or playlist. These differences in stimulus allocation and task demands are relevant when comparing gaze across experiments.

The release contains 9,729 viewing records. Of these, 777 are Task 1 records without classification responses, while 8,952 carry a real-versus-AI judgement. The latter represent 5,300 distinct verdicts because the response following a double watch is attached to both viewing records. Experiments one, two and three contribute 796, 1,588 and 2,916 distinct verdicts, respectively. Classification accuracy uses distinct verdicts, while gaze coverage retains the two watches separately.

Experiment three contributes 79,170 fixations, or 58.7% of the total, through its larger number of clip exposures and repeated viewing. This does not imply denser coverage of each individual clip. The median number of observers per clip is 20 in each of the first two experiments and eight in experiment three; the median fixation count per experiment-three clip is 222.5 across both watches. Of the 134,820 retained fixations, 130,736 fall within the displayed video. Per-clip coverage is reported within the relevant experiment, preserving the distinction between unique clips and experiment-specific temporal versions.

C VIDEO GENERATION DETAIL

The source material comprises 60 physical scenes from Physics-IQ and 20 nature scenes from Adobe Stock. Physics-IQ provides recordings from three perspectives of the same scene, , covering physical phenomena such as solid mechanics, fluid dynamics, optics, thermodynamics and magnetism. The Adobe Stock clips span nature, wildlife, food and drink, and sports. Selected clips had no generative-AI label or identifiable humans, at most two main target objects and a duration below 20 seconds.

Videos were generated through the Adobe Firefly web app using both text and image conditioning. For the physics scenes, the text prompt was the scene description supplied by Physics-IQ. In experiment one, Veo generations for the 20 physics scenes were conditioned on the supplied Physics-IQ switch frame. In experiment two, Veo generations for the 40 additional physics scenes were conditioned on the first frame. The 20 Adobe Stock scenes in experiment one used the listing description and the first frame of the source clip. Experiment-three generations also used the Physics-IQ description together with a supplied first or switch frame. These inputs specify the scene content and initial visual context; the generator determines the subsequent physical evolution.

The full dataset includes outputs from six generators. Veo 3.1 was used in experiments one and two, producing 80 clips across the physics and Adobe Stock sources. Experiment three used Firefly, Kling, Pika, Runway and Sora 2, with one generation from each model for each of the 60 physics scenes. The Firefly generation model is distinct from the Adobe Firefly interface through which the models were accessed.

All clips were processed to a common resolution of 960×720 pixels, with generator watermarks removed by cropping. Physics clips were manually trimmed to approximately five seconds. All 360 clips in experiment three, including the real references, are exactly five seconds long; Veo physics clips range from 4.96 to 5.04 seconds. The Adobe-derived block was not temporally trimmed: its generated clips are approximately eight seconds long, and its real clips range from 5.11 to 15.98 seconds. Frame rates vary across sources, including variable-frame-rate videos, so temporal alignment uses decoded frame timestamps rather than an assumed common frame rate. Scene correspondence and common output dimensions do not imply identical motion or pixelwise alignment between real and generated videos.

The released metadata records the properties of the processed clips. Complete historical generation settings, including seeds, attempt counts and output-selection logs, were not retained, and the exact crop and trim commands are unavailable. The planned real-video reconstruction script will be provided with the future public repository.

D HUMAN IN-PERSON AND ONLINE JUDGEMENT COMPARISON

We compare laboratory judgements with an unaided online classification study using 120 clips drawn from the 60 Physics-IQ scenes. The comparison includes 40 Kling videos, 40 Pika videos and 40 real videos, with the online study using the experiment-three real-video versions. The figure A.2 presents the comparison at the clip level.

Laboratory accuracy is computed from one verdict per participant and clip. In Task 3, the participant watches a clip twice before giving a single response. This response is counted once, although it is stored with both viewing records. Responses were spoken and transcribed, and participants received no trial-performance feedback. Kling and Pika judgements in this comparison come from experiment three. For real videos shared across experiments, the laboratory aggregates also include judgements from the earlier experiments that used the same video version.

The online comparison contains 9,775 unaided judgements from 604 participants. Unaided trials preceded aided trials, and only the unaided responses are included here. Each online participant judged a clip at most once, although replays were permitted. The median number of judgements per clip was 76 online and eight in the laboratory. The studies also differed in presentation: laboratory viewing used a fixed desktop display, whereas most online judgements were made on mobile devices. These differences in viewing opportunities, devices and sample sizes should be considered when comparing individual clips.

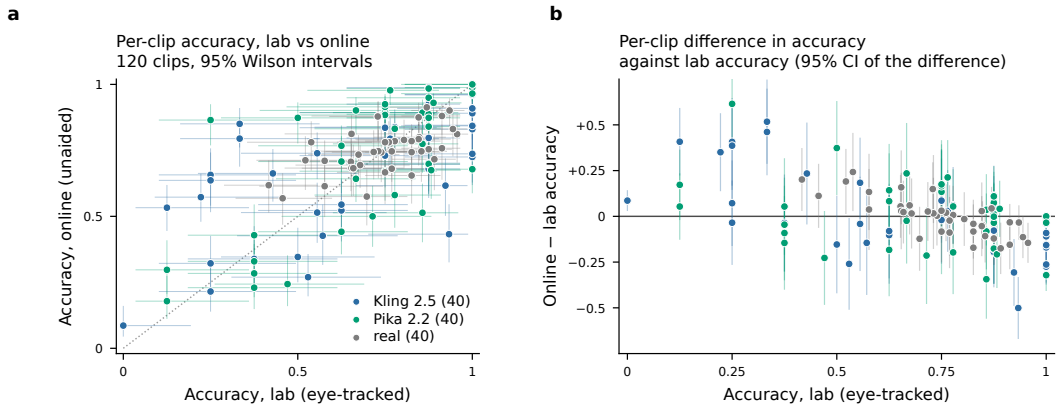


Figure A.2: **Human real-vs-AI accuracy in the eye-tracked lab sessions and in an unaided online study, per clip.** The 120 clips shown to both cohorts (60 Physics-IQ scenes; 40 Kling 2.5, 40 Pika 2.2, 40 real) were judged by 129 lab participants (2,863 judgements; 12–48 per clip, median 16) and by 604 online participants in the unaided arm (9,775 judgements; 59–126 per clip, median 76). **(a)** Online against lab accuracy for each clip; whiskers are 95% Wilson intervals and the dotted line is the identity. **(b)** Difference in accuracy (online − lab) for each clip against its lab accuracy, with the 95% confidence interval of the difference. Accuracy in the two settings is moderately correlated (Spearman $\rho = 0.60$) with no systematic shift (mean difference $+0.005$, 95% CI $[-0.028, +0.037]$); the per-clip scatter (SD 0.18) exceeds what sampling variability alone would produce ($\approx$0.13), indicating a modest setting-specific component. Lab judgement counts include both watches of a clip in the double-watch protocol, so lab intervals are slightly narrower than the number of independent viewers warrants.

Pooled accuracies were similar across the two studies. Laboratory and online accuracy were 66.0% and 66.5% for Kling, 71.3% and 72.6% for Pika, and 74.3% and 75.2% for real footage. Each pooled accuracy is the number of correct verdicts divided by the total number of verdicts for that condition. These results support broad consistency in classification performance across the two settings, although agreement on individual clips requires closer examination. Laboratory estimates are particularly sensitive to individual responses: with eight observers, one verdict changes a clip's measured accuracy by 12.5 percentage points.

For generated videos, we use detection difficulty as an operational measure of perceived realism: a clip that is more often judged real is more convincing under the viewing protocol. Accuracy on real footage instead measures correct recognition of real content. The comparison therefore provides complementary evidence about perceived realism across study settings. Participant motivation and effort were not measured, so we do not attribute differences to either.

E ANNOTATION ANOMALY TYPES AND STATISTICS

Table A.2: Anomaly types adapted from Fu et al. (2025). Impossible Physics (†) is our addition. Counts are anomaly tubes and generated clips in the release.

Anomaly type	Description	Tubes	Clips
Object Trajectory	Motion along an unnatural path.	372	239
Object Distortion	Abnormal shape deformation of an object.	214	146
Unexpected Move	Inexplicable motion of a typically static object.	122	81
Impossible Physics†	Violation of physical law not captured by a more specific anomaly type.	95	74
Redundant Object	Appearance of an unnecessary element.	88	63
Sudden Blurring	Abrupt visual degradation of a region.	51	44
Object Splitting	A single object dividing into multiple parts.	45	31
Object Disappearance	Sudden vanishing of a visible element.	40	30
Object Merging	Distinct objects fusing into one.	33	21
Background Distortion	Unrealistic warping of the background.	10	10
Total		1,070	347

Our taxonomy adapts the nine categories of DeepTraceReward and adds Impossible Physics. The table Anomaly types adapted from (Fu et al., 2025) defines the ten released categories and gives their tube and clip counts. DeepTraceReward's principal categories describe movement and appearance failures, while physical phenomena such as lighting and liquid behaviour also appear among its less frequent uncategorised cases. Our focus on realistic physical scenes motivates a dedicated category for physical-law violations not captured by a more specific existing type. The base categories take precedence: an implausible trajectory remains Object Trajectory, for example, while Impossible Physics covers residual violations such as unsupported objects or inconsistent shadows and reflections.

The annotation unit is one object or region, one dominant anomaly type and one contiguous time interval. Annotators were instructed to identify localisable anomalies visible during normal playback, using frame-by-frame inspection to confirm the event and set its boundaries. Encoding artefacts also present in the real reference and prompt mismatches alone were outside scope. Separate tracks were used for distinct objects or separable events, while overlapping descriptions of the same event were resolved through category precedence. Spatial boxes covered the affected region, and temporal boundaries marked the first and last annotated frames. Short contextual intervals could accompany instantaneous events, so tube duration may include surrounding context.

Annotations were produced with a purpose-built tool supporting per-frame box drawing, linear interpolation between annotated frames and inspection of the corresponding real clip. One annotator handled the non-Veo generations, and the other handled all Veo generations, including both Physics-IQ- and Adobe Stock-derived clips. The annotators then cross-reviewed the annotations and resolved discrepancies through discussion to reach consensus.

The Figure A.4 illustrates all of the anomaly types with example annotations. Boxes can change position and size at each frame to follow a local anomaly along a large trajectory. This retains spatial specificity that would be lost with a single stationary rectangle enclosing the whole path. For example, an Object Trajectory tube in `runway/0104_runway` has a mean frame-level box area of 0.246% of the image, whereas a stationary rectangle covering all its positions occupies 32.2%. This comparison illustrates the representation's spatial specificity; it is not a measurement of annotation accuracy or downstream performance.

The dataset contains 1,070 anomaly tubes across 347 of the 380 generated clips. Object Trajectory is the most frequent category, with 372 tubes, followed by Object Distortion with 214. Impossible Physics accounts for 95 tubes across 74 clips. Tube counts and clip prevalence measure different properties because several anomaly types may occur within one clip. Generator comparisons also depend on exposure: the five experiment-three generators share the same 60 scenes and five-second duration, whereas Veo additionally includes the longer Adobe-derived block. Comparisons of these counts should therefore account for differences in source composition.

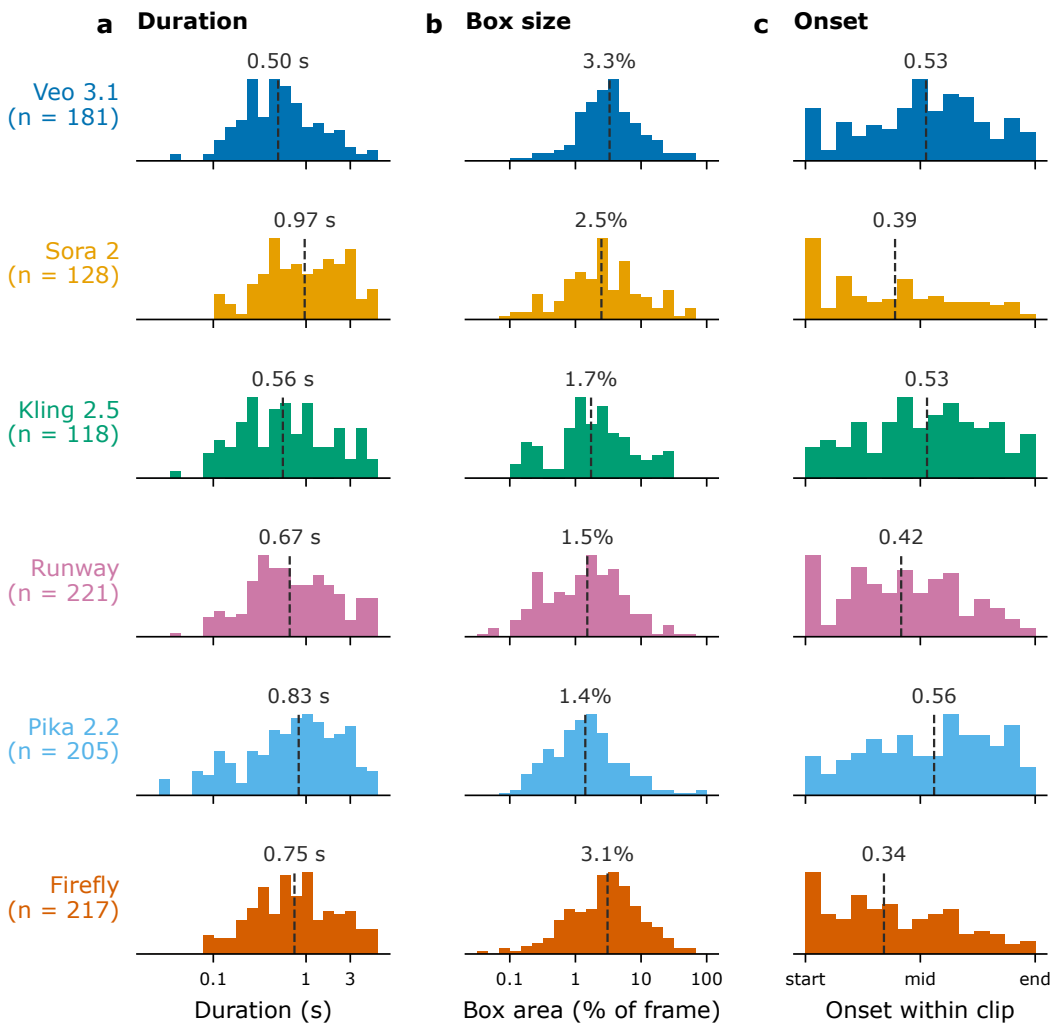


Figure A.3: **Anomaly-tube geometry by generator.** Distributions of **(a)** tube duration in seconds, **(b)** mean bounding-box area as a percentage of the frame, and **(c)** onset as a fraction of clip length, one row per generator (n = anomaly tubes; 1,070 in total). Dashed lines mark the per-generator median, printed above each panel; duration and area are on log axes. Sora 2 produces the longest anomalies (median 0.97 s vs. 0.50 s for Veo), Veo and Firefly the largest (median area ≈3% of the frame, about twice that of the other generators), and Sora 2 and Firefly the earliest-onset ones (median 0.34–0.39 of the clip vs. 0.52–0.56 for Veo, Kling and Pika).

The Figure A.3 summarises duration, spatial extent and onset. For each tube, spatial extent is the mean bounding-box area across its frames, expressed as a percentage of the video frame. We then summarise these values across tubes, giving equal weight to each tube. The per-generator medians range from 1.41% for Pika to 3.31% for Veo. Duration is the difference between the first and last annotated-frame timestamps, plus one mean frame duration to include the final frame. Median durations range from 0.54 seconds for Veo to 1.00 second for Sora 2. The onset panel expresses the start time as a fraction of the corresponding clip's duration.

Across all 1,070 tubes, onset in seconds and duration have a Spearman correlation of -0.385. This relationship is partly constrained by the finite observation window: a late-onset anomaly has less time to persist before the clip ends, and 22.7% of tubes end on the final frame. Mean tube area and duration have a weaker positive association, with Spearman correlation $+0.171$. These statistics describe the geometry of the annotated events. They do not establish that small boxes are more accurate or that a larger affected region causes an anomaly to persist longer.

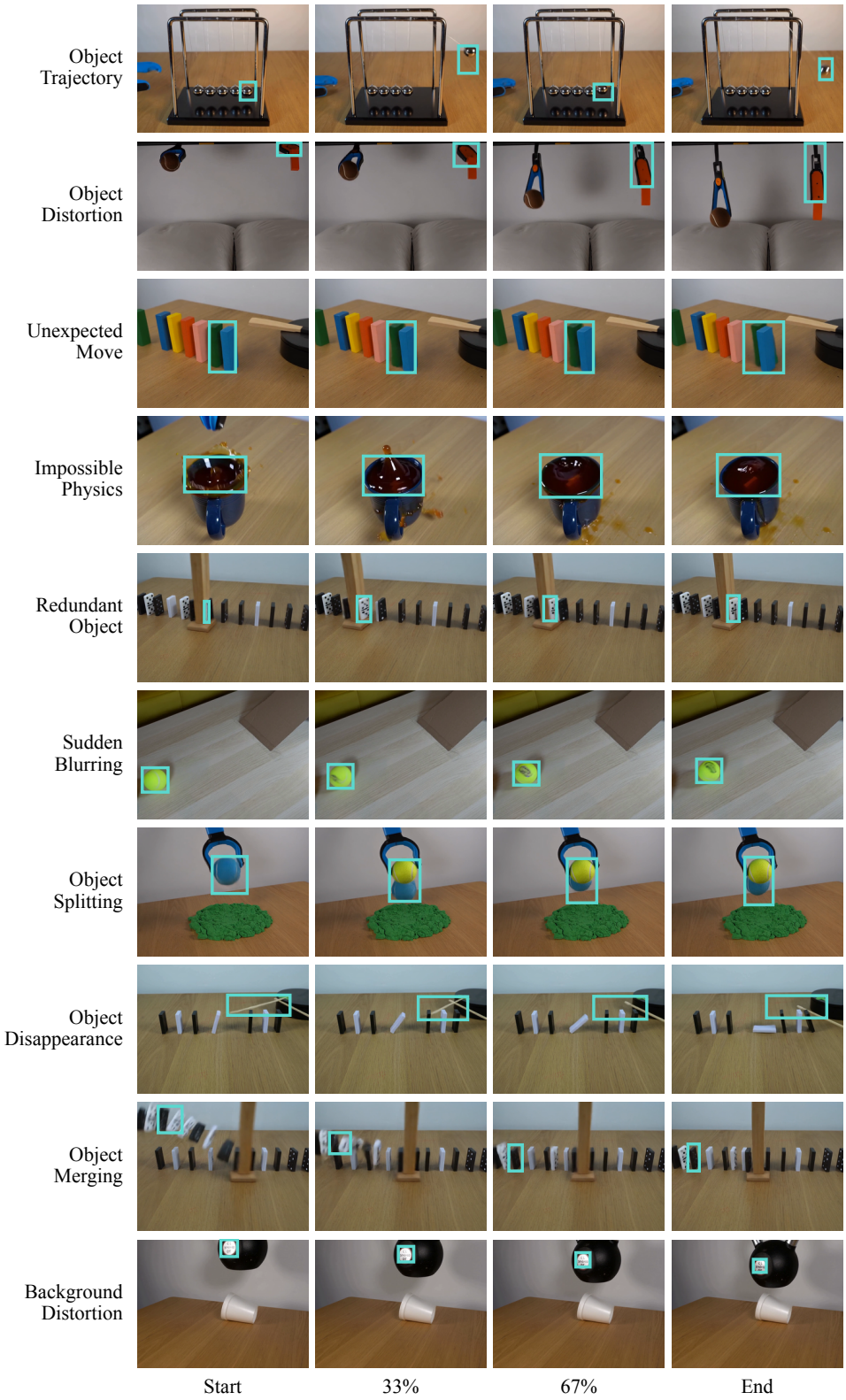


Figure A.4: **Examples of annotated anomalies**. Panels show matched spatial crops at the start, one-third, two-thirds, and end of each annotated interval.

F BENCHMARK IMPLEMENTATION DETAILS

F.1 VIDEO PREPROCESSING

Frame extraction for input into the VLMs is performed the same for all models, derived from `.mp4` files of the released GenTell-Gaze Dataset. 8 FPS was chosen uniformly as the suitable balance between meeting token budget and capturing fast object motions. The method for processing the dataset video to input frames is detailed below.

To meet the 8 FPS design, 40 frames were sampled uniformly as all clips lasted 5 seconds in duration. Targets were phase-centred across the clip duration, that is 0.0625 s, 0.1875 s, . . . , 5.0000 s. Then, the tool ffprobe was used to decode the timestamp of each frame in the clip. To account for the varying frame-rate of the clips, the nearest frame timestamp to each target was sampled, with ties going to the later frame and no frame sampled twice. The frames were then extracted using tool ffmpeg at the dropped resolution 960x720→512x384. This resolution sustains the original aspect ratio, sustains detail for recognition of small objects, and reduces image input-token cost sufficiently. Filenames, generator names, labels and timestamps are not visible to the model. This combats shortcut solutions from the generators from metadata inspection instead of reasoning on the frames themselves.

F.2 MODEL SERVING AND INFERENCE

Table A.3: **Full Model Versions** *Model Checkpoint* provides the exact model name; *Git commit SHA* is the specific commit hash on Hugging Face; *Type* highlights key the model information

Model Checkpoint	Git commit SHA	Type
Qwen/Qwen3-VL-8B-Instruct	0c351dd01ed87e9c1b53cbc748cba10e6187ff3b	8B open weights, instruct
Qwen/Qwen3-VL-32B-Instruct	0cfaf48183f594c314753d30a4c4974bc75f3ccb	32B open weights, instruct
OpenGVLab/InternVL3_5-8B	9bb6a56ad9cc69db95e2d4eeb15a52bbcac4ef79	8B open weights, instruct
zai-org/GLM-4.6V-Flash	411bb4d77144a3f03accbf4b780f5acb8b7cde4e	9B open weights, reasoning
gpt-6-astra	-	proprietary API, reasoning

The specific models are listed in Table A.3 and open-weight models were pulled from Hugging Face. The open-weight models were ran using rented compute from an array of 80+ GB GPU's on RunPod , the models themselves were ran using inference and serving engine vLLM . Pods (RunPod terminology for a GPU instance) were launched from local scripting with tasks structured by the open-source framework Harbor . There was no system prompt and no few-shot history. Each model's default chat template in vLLM was used.

For the open-weight models, a fixed set of shared inference parameters were used: zero temperature, 6,144-token output cap (2,048-token thinking cap for GLM's reasoning only). There was no temperature, top_p, nor top_k parameter choice for Astra, per OpenAI's documentation . The model was run on OpenAI's default settings with reasoning set to medium.

vLLM server command. Each open-weight model was served on a single GPU with vLLM version 0.26.0 using the command below, with the checkpoint revision pinned to a fixed Hugging Face commit, bf16 weights, a 32,768-token context, and structured outputs constrained by the xgrammar backend. The grammar was additionally configured to disallow optional whitespace between JSON tokens, so every response is emitted as compact JSON and no part of the output budget is spent on formatting.

```
vllm serve <checkpoint> --revision <sha> --generation-config vllm
--served-model-name <name>-<sha> --host 127.0.0.1 --port 8001
--tensor-parallel-size 1 --max-model-len 32768
--gpu-memory-utilization 0.9 (Qwen3-VL-32B: 0.95)
--dtype bfloat16 --hf-overrides '<json>'
({} except InternVL: {"max_dynamic_patch": 1})
--limit-mm-per-prompt '{"image": 40}' --seed 20260916
```

```
--max-num-seqs <n>
--structured-outputs-config '{"backend": "xgrammar",
"disable_any_whitespace": true}'
[--trust-remote-code] (InternVL3.5-8B)
[--reasoning-parser glm45] (GLM-4.6V-Flash, run with
VLLM_USE_V2_MODEL_RUNNER=0)
```

F.3 Model Output Schema

All models decode under a JSON-schema grammar rather than free-text generation followed by parsing. The four open-weight models are constrained by xgrammar, vLLM's structured-outputs backend. The benchmark's Pydantic output contract is compiled into a JSON Schema and, from that, into a token-level grammar before generation begins.

GPT-6 Astra, accessed through OpenAI's Responses API, is constrained the same way at the schema level using OpenAI's own Structured Outputs mode in strict mode, applying an equivalent schema with its box-coordinate tuple reshaped to satisfy OpenAI's stricter tuple syntax.

The grammar constraint is necessary but not sufficient for a valid record. Every generated response is re-validated against the same Pydantic contract after decoding, which checks constraints the grammar alone does not enforce as token masking. Example contract checks that is not done by grammar are that box keyframes are strictly increasing and that each box has positive width and height. Responses that fail at matching the JSON-scheme are recorded invalid.

Baseline JSON Output Schema. Every model returns a single JSON object in this format, and generation is constrained. The field `is_fake` gives the clip-level verdict, true for AI-generated and false for real footage. The field `anomalies` lists up to 32 localised artifacts; an empty list is valid for either verdict. Each artifact pairs one of the ten anomaly categories with a spatio-temporal tube: inclusive, one-based start and end frame indices, bounded by the 40 frames supplied, together with one or more keyframe boxes. Each box gives a frame index and integer corner coordinates $[x_1, y_1, x_2, y_2]$ on a 0–1000 grid, with the origin at the top left. Keyframes must be strictly increasing, must include both tube endpoints, and must have positive width and height; these constraints are checked by a second validation pass after generation. Any response that fails validation is recorded as invalid and excluded from scoring.

```
{
  "is_fake": true | false,
  "anomalies": [
    {
      "tube": {
        "start_frame": <integer, 1-40>,
        "end_frame": <integer, 1-40>,
        "boxes": [
          {"frame_index": <integer, 1-40>,
          "bbox": [x1, y1, x2, y2] <each coord integer, 0-1000>}
        ]
      },
      "category": <one of the ten anomaly categories>
    }
  ]
}
```

F.4 Gaze Input Details

Gaze input source. All gaze inputs are derived from the released fixation table, restricted to experiment 3 and to each participant's first viewing of a clip. This is the most task-specific gaze in the corpus: participants had been instructed to judge whether the clip was AI-generated, and had not yet seen it a second time. Only fixations that fall within the displayed video are used. Coordinates are already expressed relative to the video frame, so no further display transform is applied. Gaze

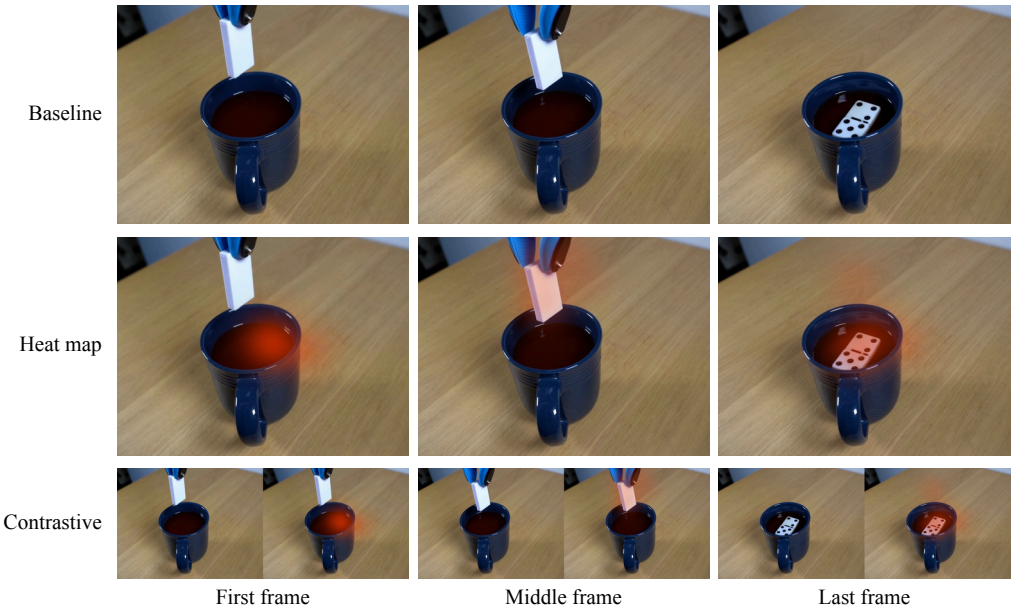


Figure A.5: **Example input frames by gaze input type**

is pooled across all first-viewing participants for each clip (6-9 participants). Gaze is aligned to the 40 frames sampled per clip. Each sampled frame is assigned a time window bounded by the midpoints between its decoded timestamp and those of its neighbours, with the first and last windows bounded by the clip. A fixation contributes to a window in proportion to its temporal overlap with that window. Figure A.5

Contrastive gaze input. Each of the 40 time steps is presented as a 1024×384 image containing two 512×384 panels: the unmodified frame on the left and a copy with a gaze heat map overlaid on the right. The heat map is a dwell-weighted Gaussian density on a 64×48 grid with σ equal to 6% of the panel width. Each fixation is weighted by its overlap with the frame's window, divided by the window length and the number of viewers of the clip. The density is normalised to a single peak per clip, so intensity encodes relative attention within a clip rather than a probability. It is rendered as an orange overlay whose opacity scales with density up to a maximum of 90%. Windows without fixations leave the right panel unchanged. The term "contrastive" refers to this side-by-side layout, not to a training objective. The prompt states that the highlights show human gaze during that interval, not annotated artifacts, and are to be treated as an attention cue rather than evidence. It also specifies that output boxes use the coordinate grid of the left panel alone and that the two panels share one frame index.

Heatmap gaze input. The heat-map input uses the same density, normalisation, colour and opacity as the right panel of the contrastive input. It is presented as a single 512×384 frame with no untouched copy. The underlying video content remains visible beneath the overlay. Output boxes use the coordinate grid of the entire image. Relative to the contrastive input, this variant removes the clean reference frame and halves the visual input per time step. It therefore tests whether gaze remains useful when it must be read directly over the content it may partially occlude.

Text gaze input. The textual input leaves all 40 frames unmodified and appends the gaze to the prompt as text, adapting the per-observer scanpath format of Gazette. The block first lists the decoded timestamp of every sampled frame. It then gives one chronologically ordered fixation sequence per anonymous viewer, with each fixation written as $(x, y, duration_ms, start_s)$. Coordinates are normalised to $[0, 1]$ with the origin at the top left, and values are written to six decimal places. Viewers are renumbered locally per clip, and fixations are never averaged across viewers or time. Every positive-duration fixation that intersects the clip is included exactly once, and fixations crossing a clip boundary are clipped to the visible interval. Relative to Gazette, the format adds

clip-relative start times, so that scanpaths can be aligned with the listed frame timestamps. It omits Gazette's fixated-object labels, which are not available for these data. It does not reproduce Gazette's training or summary-generation pipeline. This representation retains individual scanpaths that the pooled heat maps discard. The textual and visual inputs therefore differ in information content as well as encoding, so their comparison is not a pure encoding ablation. A pre-flight token count rejects any request whose prompt plus answer budget would exceed the model's context length, rather than truncating viewers or fixations.

F.5 PROMPTS

Baseline Prompt. This is the key content of the prompt to all models for the GenTell-Gaze Benchmark.

```
Is this video real or AI-generated?  The 40 images above are
ordered video frames, numbered from 1.  Return is_fake=true for
AI-generated footage and false for real camera footage.  For
example, a camera recording of a ball rolling normally is real
(is_fake=false); an AI-generated rendering of that action is
fake (is_fake=true), even if it has no visible artifact.  These
examples define the labels, not this clip's answer.  Also report
each visible generation artifact as a tube and its category.

<JSON output schema instruction>

Draw tight boxes around the visibly affected region, not an entire
object when only part of it is wrong.  Use a full-frame box only
for a genuinely global defect.

Choose one category per artifact from the descriptions below.
Examples illustrate the categories; they are not a checklist of
artifacts to invent in this clip.

<anomaly category and examples list>

Different affected objects are separate artifacts.  Do not report
the same event twice under competing category names.  When more
than one category fits, take the first that applies in this order:
an existence or count event (disappearance, splitting, merging,
or redundant object), then deformation of an object or of the
background, then motion (unexpected move or trajectory), then a
local loss of sharpness, texture, or colour stability, and only
then Impossible Physics.  Unusual subject matter, ordinary camera
effects, whole-frame stutter or judder, and compression defects
alone are not generation artifacts.  Report only visible defects
you can localise.
```

Contrastive Gaze Input Prompt. Appended block for the contrastive gaze input (duplicated frame, with and without the gaze heatmap overlay).

```
Each image is one time step with two equal-sized panels:  the
original video frame on the LEFT and a copy on the RIGHT. Coloured
highlights, when present on the right, show human gaze during
that time interval, not annotated artifacts.  Use gaze only as
an attention cue; it is not evidence that a defect exists.  Judge
the original video content, not the added highlights.  All output
boxes use the 0{1000 grid of the LEFT panel alone, never the full
two-panel image.  The two panels share one frame index; there are
not twice as many time steps.
```

Heatmap Gaze Input Prompt. Appended block for the heatmap gaze input (gaze heatmap overlaid on the frame)

```
Each image is one video frame with coloured human-gaze highlights
overlaid directly on it.  There is no separate untouched copy.
Highlights show gaze during that time interval, not annotated
artifacts.  Use gaze only as an attention cue; it is not evidence
that a defect exists.  Judge the underlying video content, not
the added highlights.  All output boxes use the 0{1000 grid of the
entire image.  Each image has one frame index.
```

Text Gaze Input Prompt. Appended block for the textual gaze input (participant scanpaths provided as text). Note also frame timestamps are provided specifically for the text gaze input.

```
The images are unchanged video frames.  The following human gaze
coordinates are normalised to 0{1, with origin at the top left, x
rightward and y downward.  Each anonymous Viewer has a separate,
chronologically ordered fixation sequence.  Tuples are (x, y,
duration_ms, start_s): duration in milliseconds, start time in
seconds from clip start.  Match fixations to the listed frame
timestamps; a fixation spans start_s through start_s + duration_ms
/ 1000.  Viewers are not averaged.  Fixations crossing the clip
boundary are clipped to the visible interval.  An empty list or
temporal gap means no valid recorded fixation, not inattention.
No fixated-object labels are supplied.  Viewer numbers are
local to this clip.  Use gaze only as an attention cue, never as
evidence of a generation artifact.  Inspect the visible content.
Output boxes still use the 0{1000 image grid.

<gaze text>
```

F.6 Evaluation Metrics Details

The definitions of the evaluation metrics reported in the GenTell-Gaze Benchmark 2 are defined here.

Evaluation Conventions. Invalid and missing outputs are reported through coverage and excluded from the metrics below, so the evaluated clips may differ between models. For classification, AI-generated is the positive class, matching the `is_fake` convention used in the baseline prompt of Section F.5. The confusions score from a True Positive is then TP_c, representing a correct model classification of a video being generated. The subscript c distinguishing classification counts from anomaly matches. Anomaly evaluation ignores category labels and is independent of the classification verdict: any listed anomalies are scored even if the clip is classified as real. AI/Real Recall, BA and F1 scores are reported as percentages, while Count MAE is reported in anomaly counts. Ratios with a zero denominator are left undefined.

Classification Metrics. The reported Balanced Accuracy (BA) is defined as the mean of the model's recall scores on either AI or real videos. The definition is such that a BA of 0.5 is equivalent the classifying each video by random chance, even considering the biased GenTell-Gaze Benchmark dataset of 300 AI compared with 60 real videos.

$$\mathrm{Real} = \frac{\mathrm{TN}_c}{\mathrm{TN}_c + \mathrm{FP}_c}, \qquad \mathrm{AI} = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c} \qquad \Rightarrow \qquad \mathrm{BA} = \frac{1}{2}\left(\mathrm{Real} + \mathrm{AI}\right). \tag{1}$$

Anomaly Count Error. Count MAE measures the mean absolute difference between the predicted and annotated numbers of anomalies per clip. For N clips with valid outputs, let $\hat{n}_i$ and n_i denote the model identified and annotated counts for clip i. Clips with no annotated anomalies, including real clips, have $n_i = 0$. Lower Count MAE indicates better agreement with the annotated counts.

$$\text{Count MAE} = \frac{1}{N}\sum_{i=1}^{N} \left|\hat{n}_i - n_i\right|. \tag{2}$$

Temporal Localisation F1. Temporal localisation measures whether predicted anomalies occur during the annotated intervals. The model output schema demands a start and end frame number for each anomaly, see Section F.3. Each of the 40 sampled frames represents a uniform temporal bin: a *predicted interval* begins at the start of its first referenced bin and ends at the end of its last. The *annotated interval* begins at the decoded timestamp of the first annotated frame and includes the full display duration of the last annotated frame. For predicted and annotated intervals P and T, temporal IoU is their intersection duration divided by their union duration, where $|\cdot|$ denotes duration.

$$\mathrm{tIoU}(P, T) = \frac{|P \cap T|}{|P \cup T|}. \tag{3}$$

Within each clip, pairs with temporal IoU of at least 0.3 are eligible for matching, irrespective of category. We find a one-to-one matching that maximises the number of eligible pairs. Matched pairs count as true positives, unmatched predictions as false positives, and unmatched annotations as false negatives. These counts are summed across all valid clips before computing F1. Predictions on clips with no annotated anomalies contribute false positives, while empty predictions on these clips leave the localisation counts unchanged. The same pooled F1 definition is used for temporal and tube localisation, with matches computed separately for each overlap measure. Precision is a measure of how often a model prediction is correct, whilst recall is a measure of how much of the ground truth is capture by the model predictions. The F1 score is an established balanced score of precision and recall.

$$\mathrm{precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \qquad \mathrm{recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{4}$$

$$F_1 = \frac{2(\mathrm{precision})(\mathrm{recall})}{\mathrm{precision} + \mathrm{recall}} = \frac{2\,\mathrm{TP}}{2\,\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}. \tag{5}$$

Tube Localisation F1. Tube localisation additionally evaluates spatial overlap. For each predicted and annotated tube pair, tube IoU is the temporal IoU of their intervals multiplied by the duration-weighted mean box IoU over their temporal intersection. Let IoU_f denote the box IoU evaluated at native frame f's decoded timestamp, and let w_f be the duration of that frame's display interval lying within both tubes' temporal intervals. Box IoU is the intersection area divided by the union area, with predicted and annotated boxes expressed on the same 0-1000 coordinate grid. The model output schema asks for a bounding box on the start and end frame, and optionally, middle frames too. We will refer to a frame which the model produces a bounding box for, a *keyframe*. Predicted box keyframes use the actual decoded timestamps of their referenced sampled frames; annotated boxes use the timestamps of their native frames. Box coordinates are linearly interpolated between keyframes and held at the nearest endpoint outside the keyframe range. For example, the last keyframe output from a model is before the models endframe for an anomaly - the keyframe's bounding box is held till the end frame in this edge case.

$$\mathrm{tubeIoU} = \mathrm{tIoU} \times \frac{\sum_f w_f \,\mathrm{IoU}_f}{\sum_f w_f}. \tag{6}$$

Tube IoU is zero when the tubes do not overlap in time or either tube has no boxes; a zero-area annotated box contributes zero box IoU. Missing boxes do not prevent temporal scoring. This definition weights box IoU by duration and differs from volumetric IoU. Tube F1 uses the matching and aggregation procedure in Equation 5, with eligible pairs satisfying a tube IoU of at least 0.3 and category labels ignored.

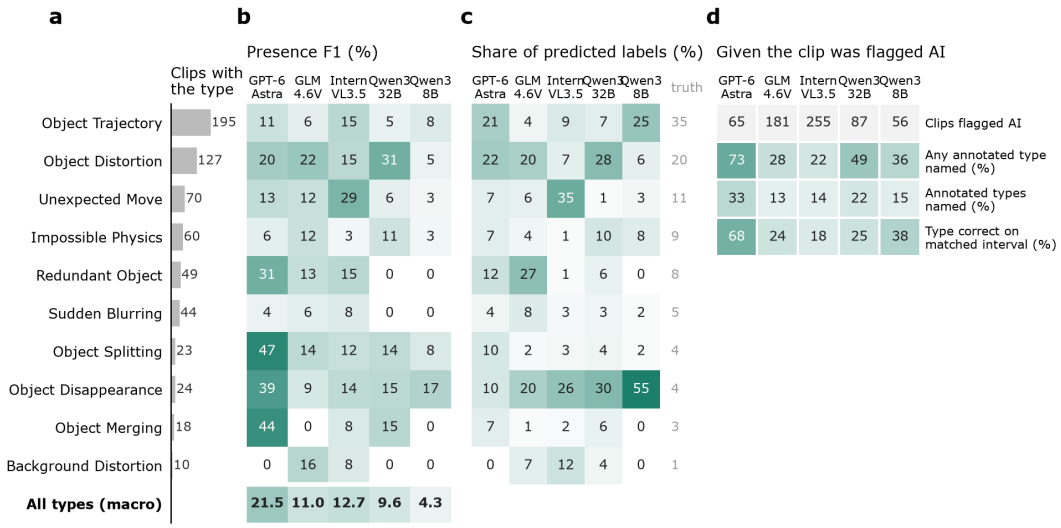


Figure A.6: **Anomaly-type recognition**, no-gaze runs on the 300 generated clips of the benchmark cohort. **(a)** Number of clips containing each of the ten anomaly types. **(b)** Clip-level presence F1 per type: whether the model names the type on the clips that contain it, irrespective of where or when. The bottom row is the macro average over the ten types. **(c)** Distribution of each model's predicted labels over the ten types, with the annotated distribution (share of all 1,070 tubes) in the right-hand column. **(d)** Type recognition on the generated clips each model flagged as AI: the number of such clips; the share of them on which at least one annotated type is named; the share of all annotated types in them that are named; and, among predicted intervals overlapping an annotated tube (IoU $\geq$ 0.3), the share that also carries the correct type (typed over untyped temporal F1).

G Additional Results

G.1 Anomaly Detection Versus Judgment

Table A.4: **A reported artifact is always accompanied by an AI verdict.** Valid predictions from every configuration in the benchmark release, pooled over the five models, four gaze inputs (none, textual, contrastive, heat map) and real and generated clips: the 360-clip main runs use the baseline wording; the three other wordings come from the 120-clip prompt-sensitivity study.

Prompt wording	Predictions	With $\geq$1 anomaly	…judged real	Judged AI, none listed
baseline	8,354	3,573	0	883
gaze as prior	1,879	979	0	157
`is_real` field	1,912	387	0	64
localise first	1,915	327	1	367
All	**14,060**	**5,266**	**1**	1,471

We investigate the link between predicting an anomaly and classifying a video as fake in Table A.4 to test model logic. Pooling valid judgements across all prompt wording variants (See Section G.6 for prompt sensitivity details), we find the reassuring result that amongst 5,266 predictions with at least one anomaly, only 1 still classified that video as real (sora2/0134_sora2 by GLM). On the other hand, there are many instances (1,471) where a clip judged to have no anomalies was still claimed generated. This was allowed in all of the prompt variants and this provides evidence of understanding the task. On the ground truth perspective, there were 18 generated clips which had no annotations, AI generated videos which an annotator found no anomaly recognisable from human observation.

G.2 Anomaly Type Categories

Ordering the rows by the most-to-least anomaly type occurrences in distinct clips provides the structure to investigate model ability to identify anomaly type. The anomaly types are input with a description and example into all model prompts. Figure A.6b explores the Presence F1 score - Presence F1 measures whether a model correctly identifies the anomaly types present in each clip, penalising missed or incorrectly reported types. It ignores timing, location and repeated occurrences of the same type. Each of the ten anomaly types receives equal weight in the final macro score.

It is interesting to observe that Astra performs well on 'count-based' anomalies (anomaly types in which the number of objects changes e.g. Object Splitting $1 \rightarrow 2$, Object Disappearance $1 \rightarrow 0$, Object Merging $2 \rightarrow 1$, Redundant Object $0 \rightarrow 1$) with an F1 score of 31%-47%.

Figure A.6c instead shows the proportion that each anomaly type takes amongst those identified by each model. Consiering the rows are ordered by ground-truth frequency, a match by the models would produce a dark-to-light gradient. It is then clear that Object Disappearance was a favourable anomaly for all models with a greater share than the ground truth (4%). In particular, Qwen-8B gave a 55% share of it's anomaly predictions to Object Disappearance alone, a heavy skew.

Separate to the anomaly type list, Figure A.6d works from the raw number of clips classified as AI (300 AI videos in total) by each model. Notice the second row, the proportion of the clips classified as AI by the model from which it also corrected labelled at least one anomaly. The worst performance was by Intern at only 22% of it's own AI classified video with a single correctly announced anomaly type. The performance for the third row degrades significantly (50% decrease) for predicted all of the anomaly types correctly. The final row considers: relative to the predicted anomalies which overlap sufficiently with a ground truth annotation, how many were correctly typed? Astra performs well here, once it's predicted anomaly overlaps with a ground truth sufficiently, it predicts the anomaly type correct 68% of the time. This implies good adherence to the anomaly type scheme. The main GenTell-Gaze Benchmark however, reports only performance metrics untyped to separate model ability to adhere to our type scheme from video reasoning and recognising an event as simply anomalous.

G.3 Comparing to human difficulty

Because every clip carries in-person judgements, model verdicts can be read against how hard each clip was for people. Figure A.7 bins the 300 generated clips by the share of viewers who judged them AI and shows, for GPT-6 Astra, the share of clips flagged in each bin together with the model's response time; Table A.5 gives the corresponding rank correlation for all five models. GPT-6 Astra's detections concentrate on the clips humans find easiest - 5% of the hardest bin against 42% of the easiest (ρ = 0.34) - and its response time rises with human accuracy (ρ = 0.42), so the model both detects more and deliberates longer where the evidence humans respond to is strongest. Splitting response time by verdict shows the rise is carried by the clips it misses: catches take a roughly constant 19 s regardless of difficulty, whereas misses lengthen from 6.4 s on the hardest clips to 10 s on the easiest, and the longest of those summaries most often appeal to a hidden mechanism to explain the evidence away. For the open models the correlation with human difficulty is weak (ρ between 0.05 and 0.16) and, where the asymptotic p is below 0.05, the scene-bootstrap intervals only just exclude zero; their verdict rates are nearly flat across human difficulty.

G.4 Astra Reasoning

The Astra model outputs it's summarised reasoning before the fixed output schema. The Table A.6 explores the observed phenomena of noticing anomalies in it's reasoning yet classifying a video as real. Manual post-processing to explore the size of this effect, was applied with increasingly strict sensitivity of whether an anomaly was reasoned about, hence classifying the video as generated. The criteria was developed from the rubric coding of the reasoning summaries from Astra such as the stated conclusion on defect finding as well as lexical flags within the reasoning summary for anomaly-noticing and language explaining away it's found anomaly. These results were derived from the full run with Astra and no gaze input. A trade-off was found for maximising the BA of video classification, performing best at 65.8% when the reasoning states a anomalous region or was unsure. If the sensitivity was maximised and any indication of anomaly-noticing language was

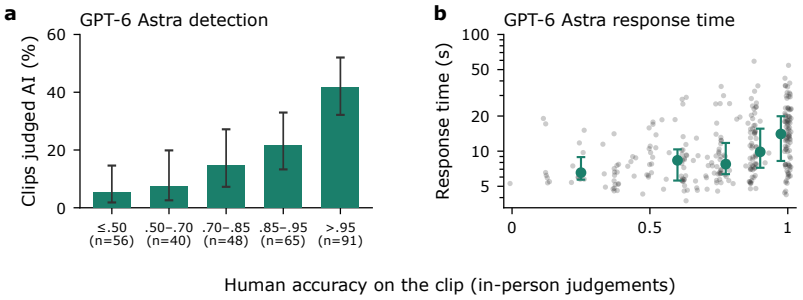


Figure A.7: **GPT-6 Astra's detections and response times against human difficulty.** The 300 generated clips of the benchmark cohort are binned by the share of in-person viewers who judged them AI-generated (median 16 judgements per clip). **(a)** Share of clips in each bin that Astra judged AI in the no-gaze run, with 95% Wilson intervals. **(b)** Astra's response time per clip (grey points) with the median and interquartile range per bin; log scale. Spearman ρ between response time and human accuracy is 0.42 ($p < 10^{-13}$).

Table A.5: **Correlation between per-clip human difficulty and model verdicts.** Spearman ρ between the share of in-person viewers who judged a generated clip AI and the model's own AI verdict, no-gaze runs on the 300 generated clips of the benchmark cohort (298 valid for InternVL3.5-8B and Qwen3-VL-8B). p: asymptotic; CI: 95 % interval from 4,000 bootstrap resamples of the 60 source scenes, which accounts for the five generated clips per scene.

Model	Spearman ρ	p	scene-bootstrap 95 % CI
GPT-6 Astra	+0.34	$< 10^{-8}$	[+0.23, +0.43]
GLM-4.6V-Flash	+0.05	0.376	[−0.04, +0.15]
InternVL3.5-8B	+0.12	0.036	[+0.03, +0.21]
Qwen3-VL-32B	+0.16	0.007	[+0.04, +0.27]
Qwen3-VL-8B	+0.14	0.018	[+0.03, +0.24]

expressed, the Real recall suffers too dramatically down 100%→40%, and BA approaches chance at 52%.

G.5 Gaze Input

All the following three investigations are derived from the data of the GenTell-Gaze Benchmark 2. We explore the distributions of the number of anomalies predicted per clip in Figure A.8 across different models and different gaze inputs. This only applied to the 300 AI-generated videos from the GenTell-Gaze Benchmark. Across all 4,792 predictions, none of the 2,188 that list an anomaly is judged real. Note the greyed bar are mistakes from the model when classifying a generated video as real, in all cases the model predicts no anomaly when classifying a video as real as expected. The Gaze input does change the number of clips classified AI with no anomaly listed (annotated above the zero bars). The gaze input formats raises it sharply for GLM and Intern models (contrastive: 106 and 107 clips against 34 and 69 without gaze), whereas the textual gaze input lowers it.

From another perspective in Figure A.9, the same above distributions are explored with the ground truth annotation counts (grey, same on all graphs) shown against the model anomaly annotations (model colour) across the gaze inputs and models. The scale is the proportion of all the generated clips (300) instead of the raw count in Figure A.8. Each panel gives the share of generated clips under- and over-counted in the number of anomalies identified by the model. Gaze input shifts the distributions modestly: the text gaze input moves Intern's mass from zero and 32 towards one and two anomalies, this is reflected in the GenTell-Gaze Benchmark where text gaze input brought Intern's best preformance across all metrics. Also, the heatmap reduces under-counting for GLM and Qwen-32B; Qwen-8B returns an empty list on 83-92% of clips under the no-gaze, textual and contrastive inputs and on 67% under the heatmap input.

Table A.6: **What GPT-6 Astra's detection would be if it had trusted what its reasoning noticed.** Each row flags, in addition to the clips the model actually judged AI, the clips whose reasoning summary meets the stated criterion; each row is a separate alternative, not a superset of the rows above it. No-gaze run on the benchmark cohort (300 generated, 60 real clips; 291 clips have a summary, the rest keep their actual verdict).

Flag the clip if judged AI, or if…	AI recall	Real recall	BA	Flagged (gen. / real)	
Actual verdict only	21.7	100.0	60.8	65	0
summary states a defect was found	25.7	100.0	62.8	77	0
summary states a defect was found or is undecided	40.0	91.7	65.8	120	5
anomaly noticed, then explained away	32.0	85.0	58.5	96	9
any anomaly-noticing language	64.0	40.0	52.0	192	36

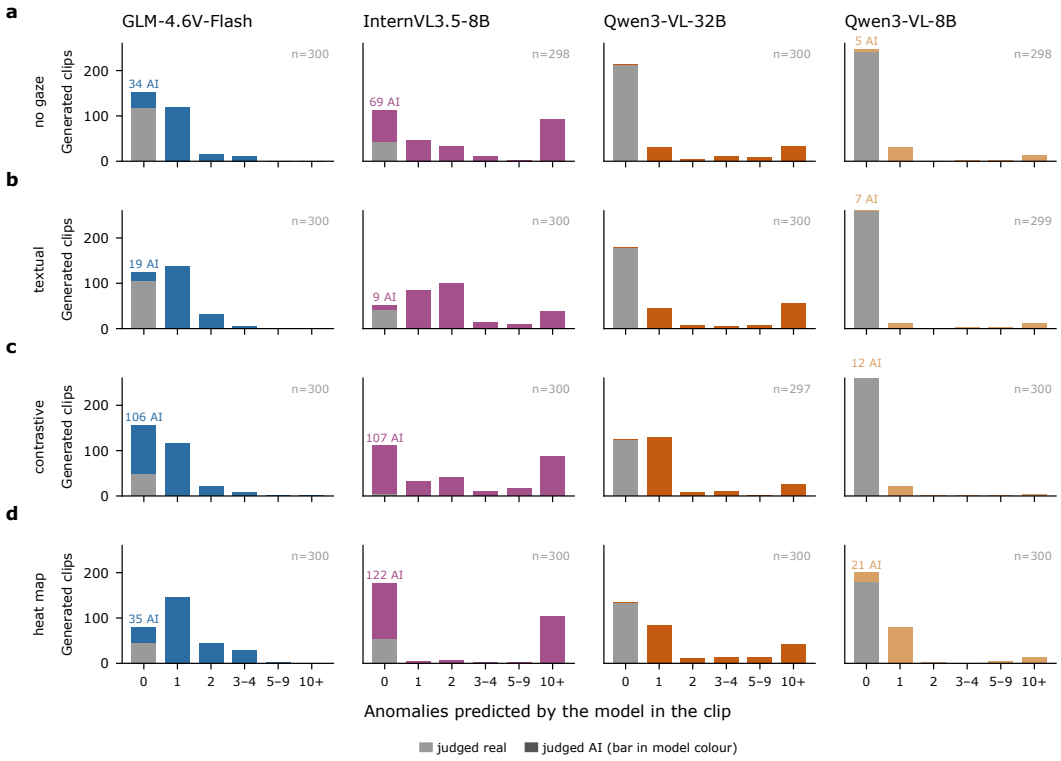


Figure A.8: **The verdict–anomaly coupling under the four gaze inputs**, open models with the baseline wording on the 300 generated clips of the benchmark cohort (n = valid predictions per panel). Rows: **(a)** no gaze; **(b)** textual fixation list; **(c)** contrastive frame pairs (each frame with and without the heat-map overlay); **(d)** heat-map overlay.

Finally Figure A.10 explores how the gaze input affects the relationship between a model judging a video as AI and the number of ground truth annotations for that video. The figure only applies to the AI-generated videos of the GenTell-Gaze Benchmark clip set, hence the vertical axis of proportion of these clips judged AI is performing best on AI recall when maximised. As with prompt wording, gaze input can shift the level of a model's curve from no gaze (grey). For example, the contrastive format raise GLM and Intern towards calling every clip AI, and the heatmap raises Qwen-8B by some 30 points - without producing a dependence on the count (Spearman ρ is -0.05-0.18 across all model-gaze pairs, against 0.42 for the human rate). Every gaze input raises the rate at which Qwen-32B judges AI whilst sustaining a similar curve to no gaze, across the number of ground truth annotations.

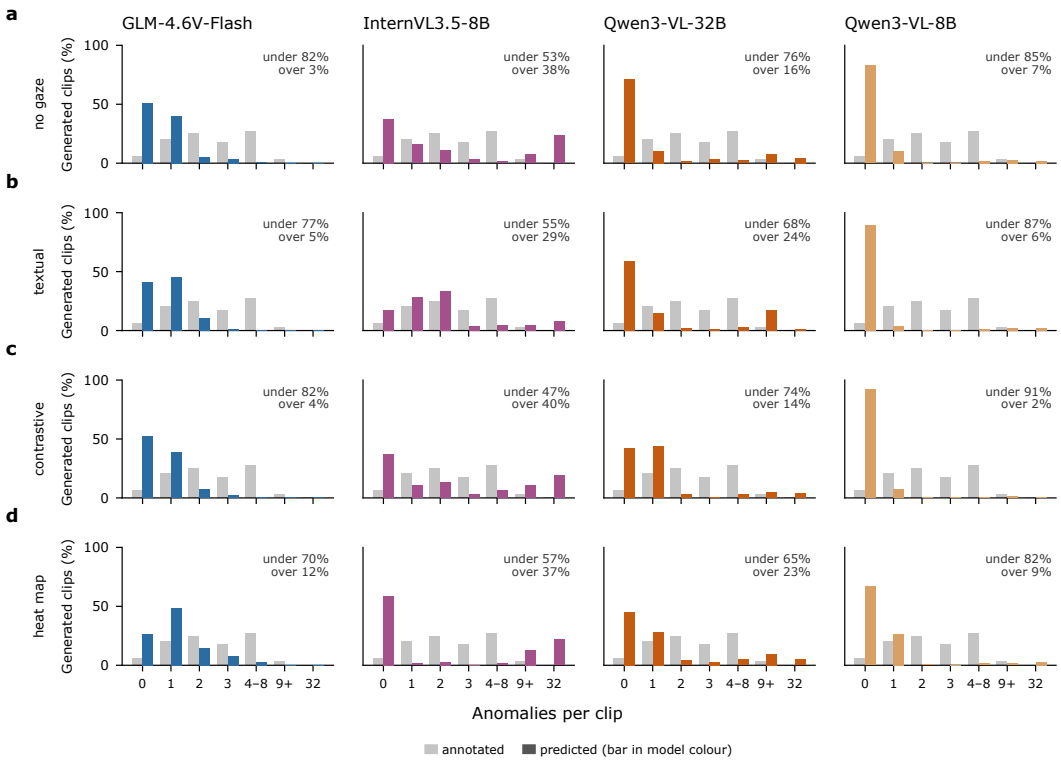


Figure A.9: **Predicted versus annotated anomaly counts under the four gaze inputs**, open models with the baseline wording on the 300 generated clips of the benchmark cohort. Rows: **(a)** no gaze; **(b)** textual; **(c)** contrastive; **(d)** heat map. Grey bars: annotated count per clip (mean 3.0, median 2); coloured bars: count returned by the model; 32 is the schema maximum. Each panel gives the share of clips under- and over-counted.

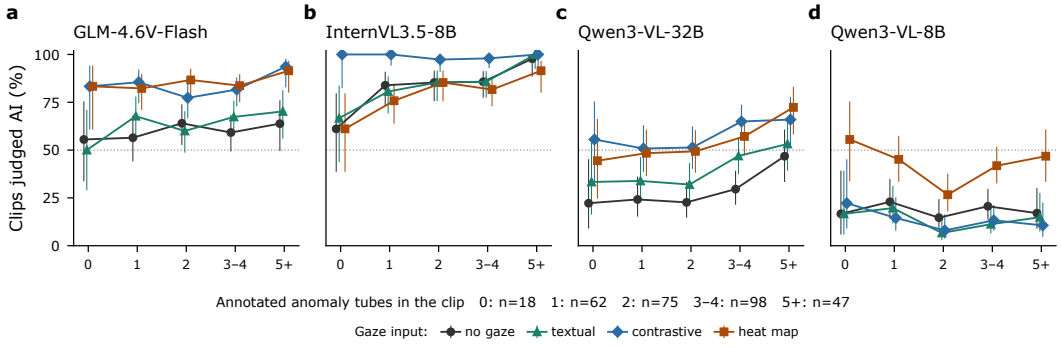


Figure A.10: **Effect of gaze input on the share of generated clips judged AI, by annotated anomaly count**, open models with the baseline wording on the 300 generated clips of the benchmark cohort (clips per bin below the axis). One panel per model, one series per gaze input. Points are offset horizontally; bars are 95 % Wilson intervals; dotted line, chance.

G.6 Prompt Sensitivity

Video subset The 120-clip subset used for the prompt-sensitivity experiments was drawn from the 360-clip benchmark cohort. Within each of the six sources (real footage and the five generators), the 60 clips were ordered by human lab accuracy, and every third clip in that order was selected, starting from the second. This gives 20 clips per source, spread evenly across the full range of human detection difficulty, so the subset preserves each source's difficulty distribution rather than favouring

easy or hard clips. The selection is fixed and identical across all models, inputs and prompt variants for the prompt sensitivity study and also used for the gaze inputs tested on Astra in Section A.8. For simplicity, we will refer to this set of videos as the subset-120, 100 generated videos and 20 real. There is also the generated only videos within the subset-120 producing the AI-subset-100.

Gaze as prior prompt. The injected change to the baseline prompt F.5 for the prompt sensitivity study on changing the strength of the gaze prior, was as follows.

```
The highlights show where people looked while judging whether
this clip is AI-generated; attention concentrates on regions
viewers found suspicious.  Treat highlighted regions as candidate
artifacts and confirm each in the original video content before
reporting it.
```

`is_real` prompt. Interested in changing the implied default of the video condition as AI or real, the following prompt sensitivity experiment was carried out. The JSON output schema, see Section F.3. Had `is_fake` swapped to `is_real` and the corresponding swap made to the baseline prompt of Section F.5.

Localise first prompt. To test asking for anomaly localisation before classifying the video, the JSON was untouched whilst the input prompt did not ask for a classification of the video as real or AI at the start, but the end of the prompt. The baseline prompt of Section F.5 was adapted accordingly:

```
The 40 images above are ordered video frames, numbered from 1.
First, report each visible generation artifact as a tube and its
category.
<baseline prompt>
Then, is this video real or AI-generated?
<baseline prompt>
```

Table A.7: **GenTell-Bench results under four prompt wordings**

Model	Prompt wording	Real	AI	BA	Count MAE ↓	Temporal F1 ↑	Tube F1 ↑
GLM-4.6V-Flash	baseline	30	61	45	2.08	17.2	0.0
	localise first	32	64	48 (+5%)	2.07 (–)	18.4 (+7%)	2.2 (–)
	`is_real` field	60	34	47 (+3%)	2.14 (+3%)	11.7 (–32%)	0.0
	gaze as prior	35	53	44 (–3%)	2.08 (–)	15.1 (–12%)	0.0
InternVL3.5-8B	baseline	29	87	58	7.98	3.4	0.0
	localise first	100	2	51 (–13%)	2.76 (–65%)	1.2 (–66%)	0.0
	`is_real` field	95	3	49 (–16%)	2.75 (–66%)	0.6 (–83%)	0.0
	gaze as prior	29	88	59 (+1%)	8.82 (+10%)	3.0 (–12%)	0.0
Qwen3-VL-32B	baseline	79	26	53	3.32	11.0	1.6
	localise first	35	50	43 (–19%)	2.35 (–29%)	0.0 (–100%)	0.0 (–100%)
	`is_real` field	75	28	52 (–2%)	3.73 (+12%)	12.4 (+13%)	1.7 (+9%)
	gaze as prior	79	27	53 (+1%)	3.54 (+7%)	10.4 (–6%)	1.1 (–30%)
Qwen3-VL-8B	baseline	80	18	49	3.34	4.3	0.4
	localise first	75	8	42 (–15%)	2.36 (–29%)	0.0 (–100%)	0.0 (–100%)
	`is_real` field	100	1	51 (+3%)	2.39 (–28%)	3.3 (–23%)	0.0 (–100%)
	gaze as prior	80	19	50 (+1%)	2.92 (–13%)	5.9 (+36%)	1.5 (+240%)

The full GenTell-Gaze Benchmark was ran on the open-weight models with the subset-120 in Table A.7 to show the performance relative to the baseline prompt. Columns are as in Table 2. Each wording changes one element of the baseline prompt: *localise first* asks for the anomaly list before the verdict; *`is_real` field* inverts the implied default video classification; *gaze as prior* rewords the gaze instruction. However, when no gaze is input, the prompt is identical to baseline, so it serves as a repeat run. Green for improvements, red for deteriorations, are shown in parenthesis relative to the baseline run. Real, AI and BA are rounded to whole numbers whilst colours follow the unrounded

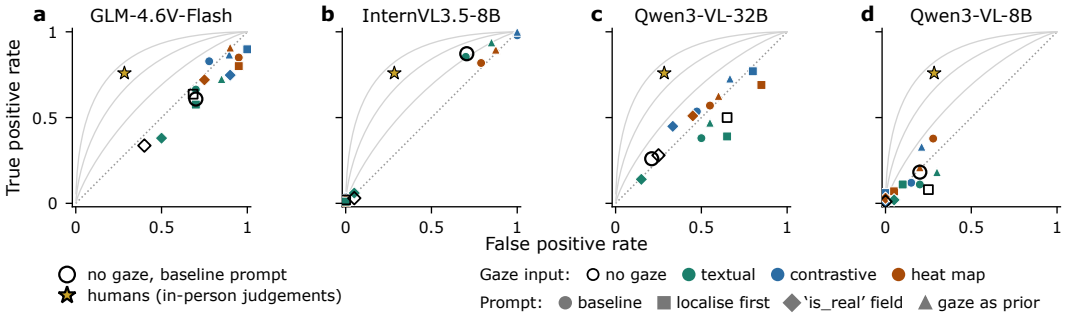


Figure A.11: **Real-vs-AI detection by prompt and gaze input, in ROC space.** One panel per open model on the 120-clip subset (20 real, 100 generated); each point is one prompt and gaze configuration. The star marks human judgements on the same clips, the dotted diagonal is chance, and grey curves join points of equal discrimination. Configurations spread along the diagonal rather than away from it.

change. The change under prompt wording is clearest in the swing of the models to favour AI or real classifications. For example, Intern has 2-3% AI recall under *localise first* and *`is_real` field* but 87-88% of the same AI recall under *baseline* and *gaze as prior*. *localise first* improved GLM scores across the board, which may align with GLM being the only reasoning model tested here.

Figure G.6 shows how the real-versus-AI judgements of the four open models change with prompt wording and gaze input. Each panel shows one model on the subset-120.

Every point is one configuration, a combination of prompt and gaze input. The point's height is the share of generated clips the model correctly labelled AI-generated (AI Recall), and its horizontal position is the share of real clips it wrongly labelled AI-generated (1 - Real Recall), so a perfect detector would sit in the top-left corner. Marker shape gives the prompt, Colour gives the gaze input. The large open circle marks the baseline configuration used in Table 2. The *gaze as prior* prompt uses the same wording as the baseline when no gaze is supplied, so that configuration is omitted, leaving 15 configurations per model. The star shows human judgements from the in-person experiments on the same clips, dominating performance on the top-left corner.

The dotted diagonal is chance: a point on it means the model labels real and generated clips as AI-generated equally often. The grey curves each join points with the same ability to tell real from generated clips. Moving along a curve only changes how readily the model answers AI-generated. Hence, moving along contours to the up-and-left corner indicates improvement.

For every model, the configurations spread along the diagonal rather than away from it. GLM and Qwen-32B range across its middle. Intern has mostly extreme top right or bottom left performance, showing extreme swings in classification bias from the gaze and prompt input pairs. Qwen3-VL-8B mostly sits near the bottom-left, labelling nearly every clip real. Changing the prompt or adding gaze therefore shifts how often each model classifies a clip as AI, but it does not bring any model meaningfully above chance.

Figure A.12 shows how each model's verdict relates to the anomalies it reports, using the AI-subset-100 without gaze input. Each bar counts clips by the number of anomalies the model listed, split into clips it judged real (grey, incorrect) and AI-generated (colour, correct). Rows show the four prompt wordings.

The coupling runs in one direction only. Of the 1,577 valid predictions, 402 list at least one anomaly, and none of these is judged real. A model can, however, judge a clip AI-generated while listing nothing. The prompt wording changes how often this happens: asking for the verdict as `is_real`, or requesting anomalies before the verdict, almost eliminates anomaly output for Intern and Qwen-8B, while under the localise first wording Qwen3-VL-32B judges 49 clips AI-generated without listing a single anomaly.

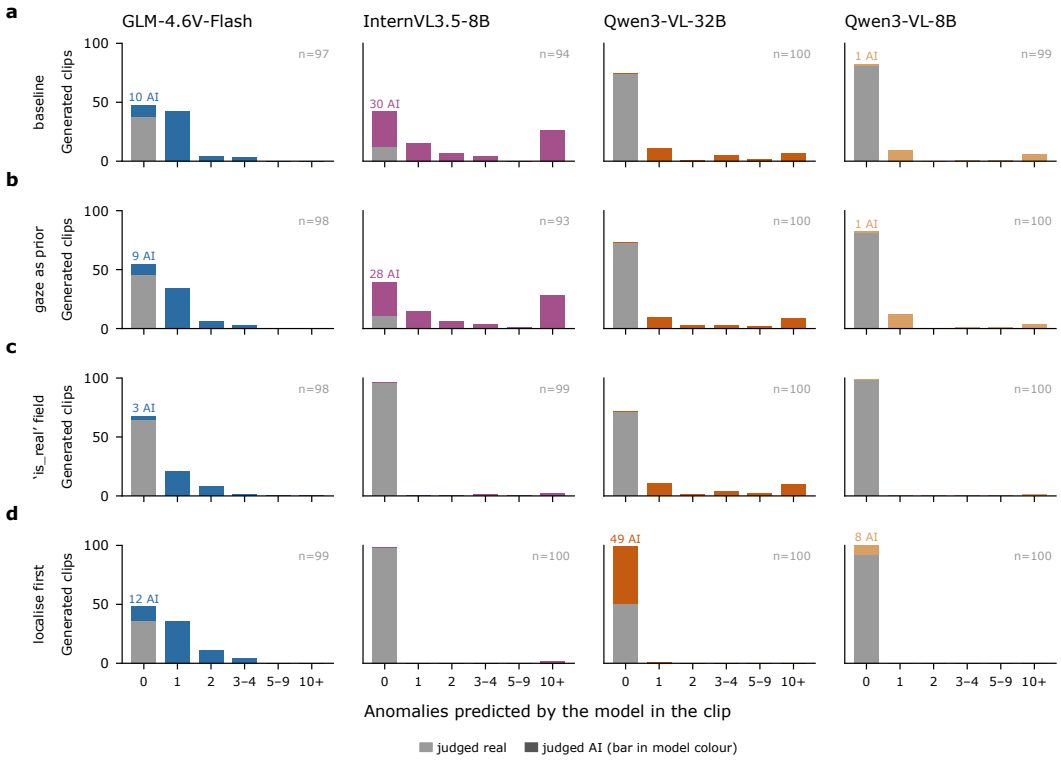


Figure A.12: **Verdict and anomaly output under the four prompt wordings.** Open models without gaze input on the 100 generated clips of the 120-clip subset; n is the number of valid predictions per panel.

Figure A.13 compares how many anomalies each model reports per clip with how many were annotated, using the 100 generated clips of the prompt-sensitivity subset without gaze input. Rows show the four prompt wordings: (a) the baseline, (b) gaze described as a prior, (c) the verdict asked as an `is_real` field, and (d) anomalies requested before the verdict. Grey bars show the annotated counts (mean 2.8, median 3 per clip) and coloured bars the counts returned by the model; 32 is the maximum the output schema allows. Each panel gives the share of clips the model under- and over-counted with respect to the ground truth quantities.

All four models under-count on most clips under every wording, but the wording changes how far this goes. Intern and Qwen-8B collapse to almost no anomaly output under both the `is_real` and localise-first wordings, under-counting 91-93% of clips. Qwen-32B collapses only under localise-first (93% under-counted); its `is_real` output (74% under-counted) is unchanged from the baseline.

InternVL3.5-8B also shows the opposite failure. Under the baseline and gaze-as-prior wordings it over-counts 35% of clips, and returns the maximum 32 anomalies on 17 and 21 clips respectively. These are clips annotated with at most five anomalies, and almost every such list repeats a single category, which suggests the model fills the list rather than reporting distinct detections.

GLM-4.6V-Flash is the most stable model across wordings (76-85% under-counted). It is also the only model that reasons before answering, which may make its output less sensitive to how the request is phrased.

Figure A.14 asks whether models, like humans, are more likely to call a generated clip AI-generated when it contains more annotated anomalies. The 100 generated clips of the prompt-sensitivity subset are grouped by their number of annotated anomaly tubes. For each group, the figure shows the share of clips judged AI-generated by each open model without gaze input and by human viewers. Each panel shows the four prompt wordings.

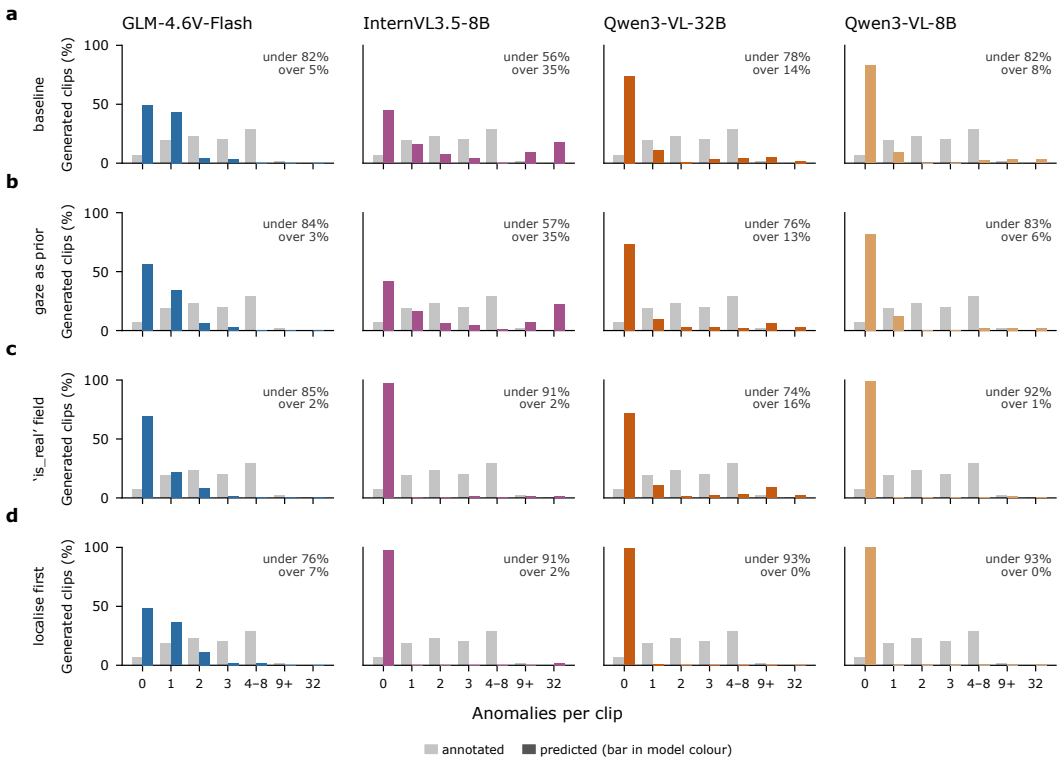


Figure A.13: **Predicted versus annotated anomaly counts under the four prompt wordings.** Open models without gaze input on the 100 generated clips of the 120-clip subset; grey bars are annotated counts (fixed), coloured bars are model counts.

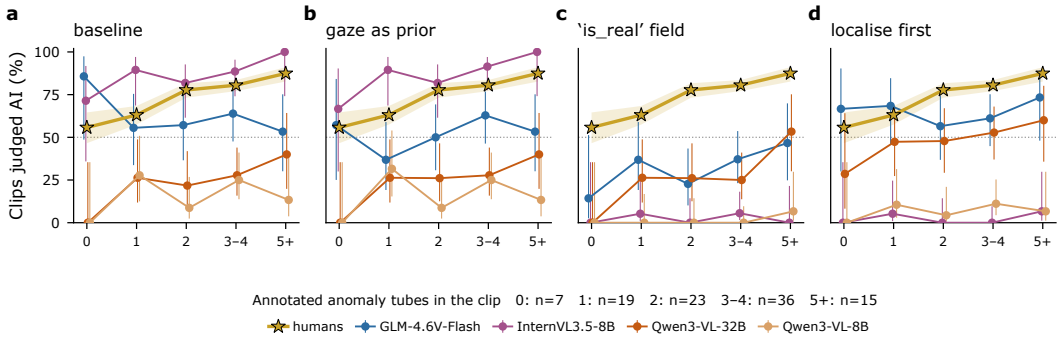


Figure A.14: **Share of generated clips judged AI-generated, by annotated anomaly count, under the four prompt wordings.** Open models without gaze input and human viewers on the 100 generated clips of the 120-clip subset; bars are 95% Wilson intervals.

Human viewers judge clips AI-generated more often as the number of annotated anomalies rises. The effect on the models is inconsistent - changing the wording raises or lowers each model's line as a whole sustaining a similar relationship between verdict and anomaly count at different levels. Measured on individual clip verdicts, the rank correlation between verdict and anomaly count lies between -0.07 and 0.18 for every model and wording, against 0.21 for human verdicts on the same clips.

Table A.8: **GPT-6 Astra with and without gaze input** on the 120-clip subset. Detection and localisation metrics as in Table 2 (temporal F1 at IoU 0.3, untyped); Reasoning section counts summaries returned, those referring to gaze, and their stance (ignored / misread as image content / used as a cue).

	Detection				Localisation			Reasoning		
Gaze	Real	AI	BA	MAE ↓	Tubes	Prec. ↑	F1 ↑	Summ.	Gaze refs	ign. / mis. / cue
none	100.0	19.0	59.5	**2.12**	26	80.8	**13.7**	74	0	0 / 0 / 0
textual	100.0	20.0	**60.0**	2.16	26	80.8	**13.7**	76	1	0 / 0 / 0
contrastive	100.0	15.0	57.5	2.20	17	**82.4**	9.4	82	2	0 / 0 / 1
heat map	100.0	18.0	59.0	2.17	22	77.3	11.2	72	13	4 / 6 / 1

Table A.8 tests whether gaze input changes GPT-6 Astra's behaviour, using the 120-clip subset (20 real, 100 generated) and the baseline prompt. The no-gaze row is the main run restricted to the same clips.

Gaze has little effect on detection. With every input, Astra labels all real clips correctly but identifies only 15–20% of generated clips, so balanced accuracy stays between 57.5% and 60.0% and count error barely moves. Localisation is similar. Astra reports few anomaly tubes (17–26 across the 100 generated clips); around 80% of them match an annotated anomaly in time, but F1 remains low (9.4–13.7) because most annotated anomalies go unreported. No gaze input improves on the no-gaze run, and the contrastive input performs worst.

Astra's reasoning summaries suggest why. They rarely refer to the supplied gaze at all: once or twice for the textual and contrastive inputs, and in 13 summaries for the heat map. Where the heat map is mentioned, the model more often ignores it or misreads the overlay as part of the scene, for example as orange lighting or colour in the video, than uses it as a cue to where artifacts may be.