VANTAGE-BENCH: EVALUATING THE INFRASTRUCTURE AI GAP IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

As Vision-Language Models (VLMs) rapidly advance toward physical deployment, the predominant focus has remained on action-oriented Embodied AI evaluated on subject-centric consumer video. This trajectory overlooks a pervasive class of Physical AI: *Infrastructure AI*, which relies on fixed-infrastructure cameras for open-loop insights like safety monitoring and operational logging. How well VLMs process dense, fixed-viewpoint visual data remains insufficiently explored. We introduce **VANTAGE-Bench**, an evaluation benchmark that measures this "Infrastructure AI Gap," distinguished by four features: 1) **Domain Relevance**, spanning three operational environments (Warehouse, Transportation, and Smart Spaces); 2) **Modality Breadth**, unifying image and video evaluation to probe semantic, spatial, temporal, and spatio-temporal capabilities; 3) **Task Format Diversity**, moving beyond multiple-choice to eight task formulations spanning discriminative QA, generative dense captioning, and spatio-temporal grounding; and 4) **Evaluation Novelty**, a single-pass trajectory-generation protocol for Single Object Tracking and, to our knowledge, the first such evaluation on fixed-camera infrastructure video, scored against specialist trackers. Annotation spans three regimes over 3,346 expert-annotated media assets: video tasks (3,342 annotations over 854 videos), image grounding (4,281 over 1,864 images), and dense detection (27,404 boxes over 628 images).

Evaluating 18 models zero-shot, we find the shortfall relative to consumer-centric benchmarks is concentrated rather than general. Event verification and temporal localization fall 11.7 to 23.5 points at all three scales we can match, while video question answering stays within 5.3 points of VideoMME and 2D spatial pointing is flat or better against BLINK. The temporal pillar is the weakest: GPT-6 Astra is the only model of the eighteen to clear 55.71 mIoU on temporal localization or 37.28 SODA_c on dense video captioning, leading the field by 9.4 mIoU on the former and 0.9 SODA_c on the latter. On tracking, the strongest frontier model matches purpose-built specialist trackers at every horizon we test, while weaker open-weight models fall to a static-box baseline. Open-weight models lead 2D object localization outright, so neither parameter count nor proprietary access accounts for the pattern. The dataset, evaluation harness, and public leaderboard are publicly released; the harness is available at an anonymized repository, and the dataset and leaderboard URLs are withheld for double-blind review.

1 INTRODUCTION

As VLMs rapidly advance toward physical deployment, the prevailing research trajectory has heavily indexed on action-oriented Embodied AI and evaluations grounded in subject-centric consumer video (e.g., egocentric recordings, broadcast media). This focus overlooks a pervasive and foundational class of physical AI: *Infrastructure AI*. Relying on fixed-infrastructure cameras (e.g., CCTV, elevated sensors), Infrastructure AI provides open-loop insights for safety monitoring, operational logging, and large-scale spatial reasoning. Because frontier models are predominantly trained on internet-crawled data, they are biased toward standard photographic perspectives. Consequently, their capability to process the dense, oblique, and multi-agent visual data inherent to fixed-viewpoint environments remains insufficiently evaluated, creating a severe implicit domain shift.

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.

Under these conditions, the "internet-video prior" fails, and models must rely on pure spatial-temporal logic. A model can achieve state-of-the-art scores on internet video and still perform poorly in the environments where these systems are actually deployed. We term this the *Infrastructure AI Gap*: the discrepancy between a model's performance on general video benchmarks and its reliability as an insight-generating system in dense, fixed-camera infrastructure environments. Closing this gap requires evaluation frameworks designed natively for operational reality. While comprehensive suites like VideoMME [1] and MVBench [2] evaluate multiple reasoning dimensions, they ultimately reduce evaluation to a single format: MCQ.

To systematically measure and close this capability deficit, we introduce **VANTAGE-Bench** (**V**ideo **An**alysis **T**asks **A**cross **G**eneralized **E**nvironments). VANTAGE-Bench is the first multi-task benchmark specifically curated for Infrastructure AI, moving beyond single-format MCQ to evaluate eight distinct task formulations across four pillars of operational visual intelligence: Semantic, Spatial, Temporal, and Spatio-Temporal Understanding.

Our main contributions are:

1. **VANTAGE-Bench:** To our knowledge the first multi-task benchmark curated for Infrastructure AI, spanning three deployment domains — Warehouse, Transportation, and Smart Spaces — and addressing the structural distribution shift from egocentric and consumer video. Eight tasks over four reasoning pillars on 3,346 fixed-viewpoint media assets, annotated by trained professionals rather than crowdworkers, under per-task rulebooks with 100% second-expert review.
2. **Format-Diverse Evaluation:** Five of the eight tasks require the model to emit bounding boxes, temporal boundaries, or a full trajectory with no candidate set supplied, removing the implicit answer prior that MCQ evaluation provides; only two are multiple-choice and one binary.
3. **A Tracking Protocol for VLMs:** A single-pass trajectory-generation protocol for Single Object Tracking and, to our knowledge, the first such evaluation on fixed-camera infrastructure video, reported against specialist-tracker, static-box, and random anchors on identical tracklets and frames.
4. **The Temporal Pillar Is Where Capability Is Still Being Acquired:** At matched scale, physical-AI post-training raises 2D Object Localization by 26.1 and 24.0 points while Temporal Localization moves 3.0 to 3.7. Only GPT-6 Astra clears 55.71 mIoU on Temporal Localization, at 65.10, yet it leads Dense Video Captioning by only 0.89 points.
5. **Diagnosis Rather Than Ranking:** The deficit survives a prompt- and metric-matched control against Charades-STA and appears *within* a single task, where dense-captioning BERTScore-F1 spans 54.2 to 63.9 while temporal IoU-F1 spans 33.2 to 42.7 on the same items. Freeze rate further separates two tracking failure modes that Success AUC conflates.

2 Related Work

2.1 General Video Benchmarks and the MCQ Bottleneck

The dominant paradigm in VLM evaluation relies on general-purpose, multi-task benchmarks such as VideoMME [1], MVBench [2], and EgoSchema [3]. While these comprehensive suites have driven substantial progress in general video understanding, their relevance to operational deployment is limited by two factors. First, they rely on a "cinematic prior"—utilizing web-sourced, broadcast, or egocentric footage characterized by high-velocity optical flow, human-centric framing, and edited temporal structures. Second, they suffer from the "MCQ Bottleneck." By reducing complex spatio-temporal reasoning to MCQ, these benchmarks provide the correct answer as an implicit prior. This format does not test the open-ended generative capabilities—such as producing precise bounding box coordinates or explicit temporal boundaries—that autonomous Infrastructure AI requires to generate reliable downstream insights.

Recent VLM benchmarks. A growing line of work probes specific capabilities through purpose-built evaluations: visual temporal reasoning [46], spatial perspective and rotation [47], and long-video understanding across timescales [48]. VANTAGE-Bench differs in setting rather than ambition. It evaluates fixed-infrastructure viewpoints, which these suites do not cover, and it reports single object tracking as a generative trajectory task across open-weight and proprietary models, which to our knowledge no prior VLM benchmark quantifies.

2.2 Embodied AI Datasets vs. Classical Surveillance

The physical AI community has developed extensive datasets for real-world deployment, but these efforts remain fragmented across two divides. First is the operational divide. Datasets like KITTI [5], nuScenes [6], and Waymo [7] are optimized exclusively for Embodied AI—featuring moving cameras, low-latency requirements, and closed-loop perception. On the other side is classical surveillance, where datasets like UCF-Crime [8] and VIRAT [9] capture the correct fixed-infrastructure perspective, but rely on classical bounding boxes or binary anomaly labels without a natural language interface. Second is the domain divide. Existing enterprise datasets are strictly siloed—focusing solely on either transportation (e.g., AI City Challenge [10]), generic crime (UCF-Crime), or robotic manipulation in logistics. VANTAGE-Bench bridges both divides simultaneously. It pairs the fixed-infrastructure perspective of classical surveillance corpora with a natural-language interface, and unifies evaluation across three distinct deployment domains: Transportation, Warehouse, and Smart Spaces. This multi-domain approach ensures models are evaluated on generalized physical reasoning rather than overfitted domain-specific priors.

2.3 Spatial and Temporal Grounding in Idealized Settings

Grounding language to spatial and temporal coordinates is a foundational requirement for physical AI, yet existing benchmarks evaluate these capabilities under highly idealized conditions. Spatial grounding and referring expression datasets like RefCOCO [13] and BLINK utilize natural indoor and outdoor photography featuring well-lit, centered, and distinct objects. This fails to evaluate the dense, multi-instance disambiguation required when monitoring infrastructure from an elevated, oblique perspective (e.g., distinguishing between dozens of identical pallets or vehicles). Similarly, temporal grounding benchmarks like Charades-STA [11] and ActivityNet Captions [12] utilize scripted, continuous human actions where the event of interest dominates the video timeline. In contrast, operational video is characterized by long quiescent periods. VANTAGE-Bench forces models to search through extended periods of inactivity to localize brief, sparse, safety-critical events, which tests temporal attention and contextual recall more directly.

2.4 Spatio-Temporal Tracking and Grounding

Our Single Object Tracking track sits between four lines of work. Classical benchmarks (LaSOT [14], GOT-10k [15], MOTChallenge [16]) evaluate dedicated pipelines that maintain explicit internal state, and we use four such specialist trackers—SAM 3 [25], SAMURAI [26], SUTrack [27], and MCITrack [28]—as anchors on our own annotations (Appendix D.4). TNL2K [17] evaluates iterative language-conditioned tracking by specialist architectures. VidSTG [18] and HC-STVG [19] score joint spatio-temporal grounding with a single number that conflates language grounding, spatial localization, and temporal localization; our pillar design separates exactly these three. Elysium [20] and Merlin [21] are purpose-trained tracking MLLMs, VideoChat-R1 [23] and R1-Track [24] address zero-shot box-initialized tracking, and MMT-Bench [22] evaluates tracking as multiple-choice box selection. Our protocol differs in kind: the model receives the target box overlaid on the first frame and emits a continuous trajectory in a single pass, without rolling memory updates.

3 Dataset and Benchmark Construction

To construct a benchmark capable of testing the Infrastructure AI gap, we curated a multi-modal dataset combining expert-annotated real-world infrastructure footage with a limited set of high-fidelity simulated environments. The dataset is human-labeled. Automated procedures are used to structure and scale one task format, 2D Spatial Pointing, where the question and its correct answer are both derived deterministically from human-labeled box coordinates (Section 3.3). VANTAGE-Bench comprises 3,346 media assets across eight tasks (see Figure 1); Figure 2 shows a worked example of each. The video tasks and 2D Object Localization draw on fixed-infrastructure cameras: static, elevated, wide-angle viewpoints at mounting heights of 8 to 20 feet and viewing angles of 30 to 60 degrees from vertical, including low-light, night, and rain conditions. These are highway ITS and warehouse installations. The two remaining image tracks sample different geometries: 2D Referring Expressions uses oblique aerial views from RefDrone [29], and 2D Spatial Pointing is derived from object-localization boxes over vehicle-mounted footage. Both are single-frame tasks, so platform

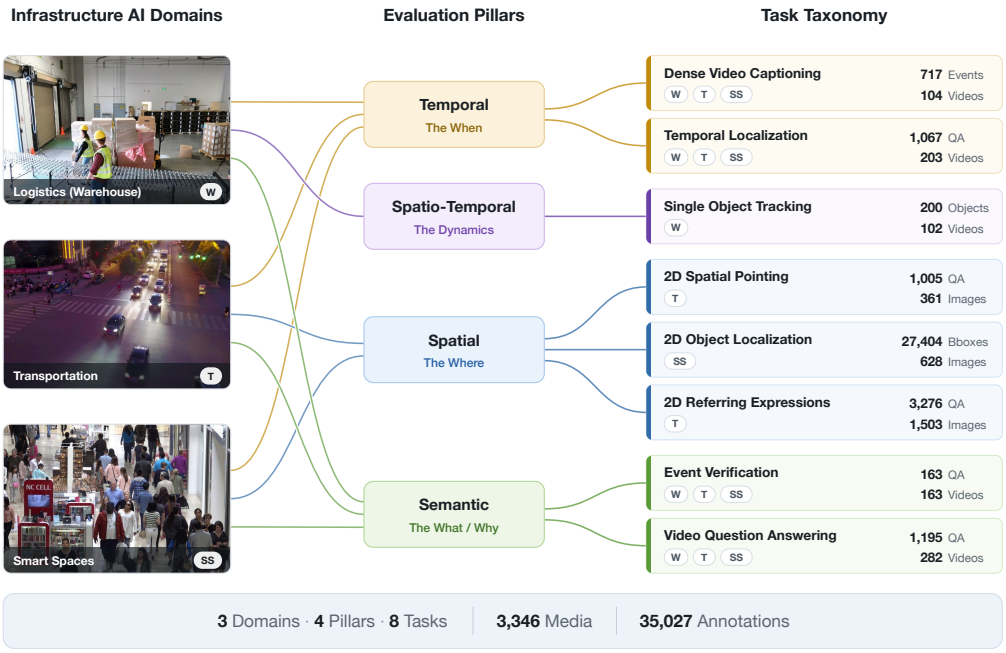


Figure 1: Taxonomy and distribution of VANTAGE-Bench. The benchmark spans three operational domains (W: Warehouse, T: Transportation, and SS: Smart Spaces), evaluating eight distinct tasks across Semantic, Spatial, Temporal, and Spatio-Temporal pillars. In total, the dataset comprises 3,346 media assets across three annotation regimes: 3,342 video-task annotations, 4,281 image-grounding annotations, and 27,404 detection boxes.

motion never enters the evaluation; what varies across the image tracks is viewpoint geometry, which is the property we intend to sample. Single Object Tracking uses synthetic elevated-fixed warehouse sequences from PhysicalAI-SmartSpaces [30]. To support diverse physical reasoning, approximately 20% of the Video Question Answering and Temporal Localization splits feature high-fidelity synthetic simulations. VANTAGE-Bench is publicly released under an **evaluation-only data license**, which strictly restricts dataset usage to evaluation and benchmarking purposes. The dataset URL is withheld for double-blind review.

3.1 Task Taxonomy

We organize VANTAGE-Bench around four reasoning pillars covering the functional capabilities reliable Infrastructure AI requires, with formal definitions and metrics in Appendix B. **Semantic Understanding** evaluates causal and operational reasoning through Event Verification, which checks operational hypotheses against visual evidence, and Video Question Answering over untrimmed sequences. **Spatial Understanding** evaluates geometric grounding through 2D Referring Expressions, which requires dense semantic disambiguation, 2D Spatial Pointing, which requires coordinate selection from positional prompts, and 2D Object Localization. **Temporal Understanding** covers Temporal Localization, requiring exact start and end timestamps, and Dense Video Captioning, which must both localize and describe chronological events. **Spatio-Temporal Understanding** consists of Single Object Tracking, which requires predicting a target's coordinate trajectory across a full sequence and so tests the maintenance of spatial context over time.

3.2 Source Data and Annotation

The core footage was curated by the authors to capture fixed-camera infrastructure perspectives that are hard to source. Real-world footage originates from two U.S. municipalities, supplemented by publicly sourced and licensed European footage; consent provenance and the PII obfuscation

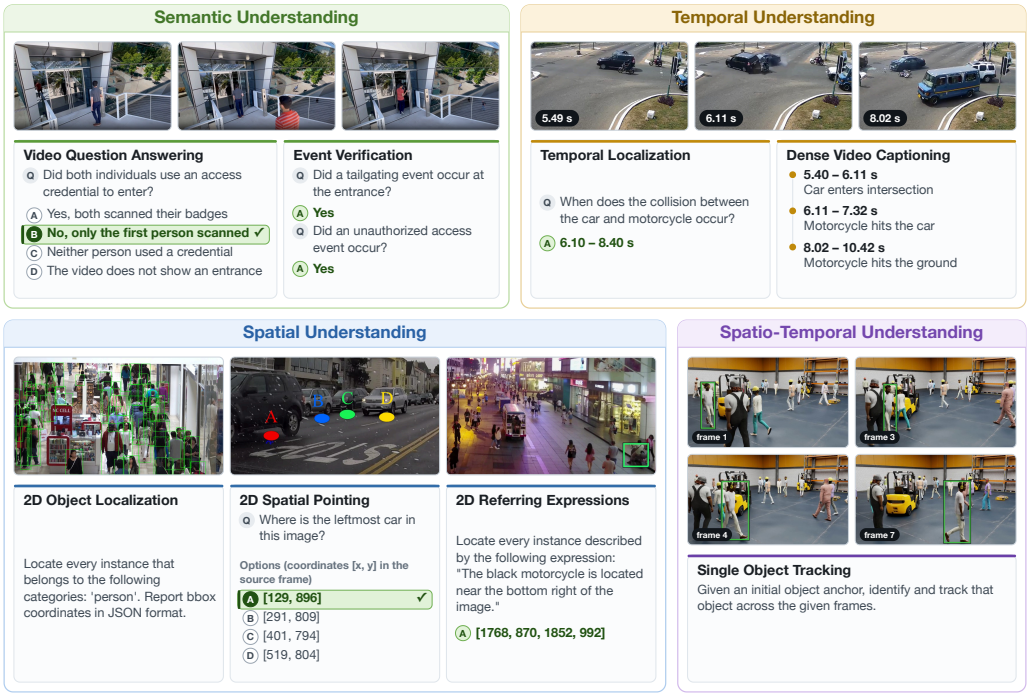


Figure 2: Visual examples of the eight evaluation tasks in VANTAGE-Bench, categorized by their primary reasoning pillar: Semantic, Spatial, Temporal, and Spatio-Temporal Understanding.

pipeline are detailed in Appendix E. Annotations were produced by trained professionals rather than crowdworkers, using domain-specific rulebooks stabilized over up to two months, with only the top 10% of the annotator pool working on benchmark data. Every annotation undergoes second-expert review, with a measured 98% first-pass acceptance and 99.43% acceptance after revision, and a third-tier expert audits a random 10% subset.

3.3 Programmatic and Adapted Tracks

2D Spatial Pointing is generated by the Pointing Data Engine, which converts human-labeled object-localization boxes over vehicle-mounted footage into multiple choice: both the question and its correct answer derive deterministically from KITTI-style box geometry, and a VLM rewrites only the question text, never which option is correct. Same-class distractors keep the choice difficult. Two further tracks adapt public sources. 2D Referring Expressions uses RefDrone [29] for its dense, oblique aerial views; those queries are human-authored, and the rewriting above applies only to Pointing. Single Object Tracking adapts synthetic multi-camera warehouse sequences from PhysicalAI-SmartSpaces [30], reformatted into our single-pass $[x, y, w, h]$ protocol. Appendix C gives filtering, reformatting, and sampling details.

4 Evaluation Framework and Empirical Results

To quantify the Infrastructure AI capability gap, we evaluate 18 VLMs on VANTAGE-Bench across the four pillars defined in Section 3.1.

4.1 Experimental Setup

We evaluate 18 models spanning four vendors and 2B to frontier scale. Twelve are open-weight: Qwen3-VL at 2B, 8B, and 32B [37]; Qwen3.5 at 9B and 27B [44]; Cosmos-Reason2 at 2B, 8B, and 32B [43]; Cosmos3 Edge, Nano, and Super [38]; and Gemma-4-E2B [39]. Six are proprietary:

GPT-6 Astra [40], GPT-5.6 Sol [41], Gemini 3.6 Flash, Gemini 3.5 Flash-Lite, Gemini 3.1 Pro, and Gemini 3.1 Flash-Lite [42]. Four of the open-weight families provide a scaling series, which we use in Section 4.4. To test out-of-the-box physical reasoning and eliminate fine-tuning biases, all models were evaluated zero-shot. Inference and metric computation use an extended VLMEvalKit harness [31]; metrics are defined in Appendix B.

4.2 Main Results

We analyze model performance across the four pillars of operational visual intelligence (Table 1). Performance is uneven across pillars rather than uniformly low. Absolute scores are not directly comparable across tasks, since the metrics differ in what a perfect score demands; Section 4.3 therefore reads each task against an external reference measured the same way.

		Semantic		*Spatial*			*Temporal*		*Sp.-Temp.*	
Model	**Params**	**VQA** (Acc.)	**EV** (M-F1)	**Point.** (Acc.)	**RefEx** (mIoU)	**Obj Loc.** (F1@.5)	**Temp Loc.** (mIoU)	**DVC** ($SODA_c$)	**SOT** (AUC)	**Overall**
Open-weight models										
Qwen3-VL	2B	63.85	44.73	53.13	65.12	47.42	35.21	19.90	30.38	41.86
Qwen3-VL	8B	66.44	59.39	68.56	73.29	59.93	44.32	29.64	33.12	50.07
Qwen3-VL	32B	71.30	60.04	75.62	72.39	72.67	46.80	29.40	44.19	55.38
Qwen3.5	9B	71.63	46.78	71.74	77.16	76.87	42.53	27.88	23.52	48.30
Qwen3.5	27B	67.95	55.34	75.32	76.35	85.46	36.88	26.97	24.40	49.25
Cosmos-Reason2	2B	64.69	55.27	59.70	56.33	73.56	38.95	28.52	26.85	45.94
Cosmos-Reason2	8B	67.95	64.09	68.60	70.26	83.88	47.30	32.50	37.69	54.46
Cosmos-Reason2	32B	70.29	73.58	74.03	50.04	4.35†	50.47	30.04	33.12	47.03†
Cosmos3-Edge	4B	62.68	63.36	67.76	69.27	79.02	39.51	16.64	24.65	46.94
Cosmos3-Nano	16B	68.95	68.88	74.83	75.57	74.11	48.04	31.42	59.21	60.67
Cosmos3-Super	64B	69.46	71.28	72.94	76.23	**86.97**	51.90	29.54	64.66	63.62
Gemma-4-E2B	2B	46.19	27.65	31.74	5.52	45.40	9.02	9.14	11.58	21.28
Proprietary models										
Gemini 3.1 Flash-Lite	—	68.03	63.78	63.68	52.70	71.74	37.72	32.54	46.68	52.61
Gemini 3.1 Pro	—	71.46	68.57	72.04	67.17	77.21	45.69	35.55	67.88	62.66
Gemini 3.5 Flash-Lite	—	65.61	70.41	43.08	74.46	65.73	24.00	32.29	55.77	53.25
Gemini 3.6 Flash	—	76.82	82.00	79.20	75.78	81.57	51.50	36.15	75.99	69.52
GPT-5.6 Sol	—	78.08	76.14	81.90	69.44	69.30	55.71	37.28	75.02	68.04
GPT-6 Astra	—	**83.26**	**86.90**	**86.97**	**80.18**	82.20	**65.10**	**38.17**	**81.20**	**75.26**

Table 1: VANTAGE-Bench main results. Zero-shot performance of 18 models on the eight tasks, grouped by pillar; all metrics scaled 0–100. *Overall* is the macro-average of the four pillar scores; bold denotes the best per column. Pillar scores are in Table 11, item-level 95% bootstrap intervals in Table 13. †Output-format regression on Object Localization for this checkpoint, reported unmodified; it propagates into the aggregates that contain it.

Semantic Understanding. Semantic understanding is the strongest pillar at the top of the table, though spatial grounding has the higher aggregate across the full model set. GPT-6 Astra leads both, at 86.90 Macro F1 on Event Verification and 83.26 on Video Question Answering, with Gemini 3.6 Flash next at 82.00 and Cosmos-Reason2-32B leading the open-weight field at 73.58. Event Verification separates models far more sharply than VQA, spanning 27.65 to 86.90 against 46.19 to 83.26. Smaller models fail asymmetrically rather than uniformly: Qwen3-VL-8B misses true events while Cosmos-Reason2-8B accepts ones that did not occur, so Macro F1 conceals two distinct behaviours. We term this **causal blindness** and decompose it in Appendix D. Event Verification is also one of the smallest splits in the suite ($n = 163$, 104 positive), where a random predictor scores 48.95 Macro F1 and an all-negative predictor 26.58. Four models are statistically indistinguishable from that random floor, and Gemma-4-E2B sits below it at 27.65, essentially the all-negative value. The two multiple-choice tasks, Video Question Answering and 2D Spatial Pointing, have a 25.0 chance floor that every model clears. Metric definitions are in Appendix B.

Spatial Understanding. The spatial pillar is where open-weight models remain competitive. GPT-6 Astra leads 2D Spatial Pointing at 86.97 and 2D Referring Expressions at 80.18, but 2D Object Localization is the one task in the suite it does not lead: Cosmos3-Super reaches 86.97 against Astra's 82.20, with Qwen3.5-27B close behind at 85.46. Dense single-frame detection is therefore the capability where open-weight models still hold the top position.

Temporal Understanding. The temporal pillar has the lowest aggregate of the four across the full model set. GPT-6 Astra is the only model to clear 55.71 mIoU on Temporal Localization, at 65.10; the best open-weight score is 51.90 mIoU. Its margins over the next model are 9.39 mIoU on Temporal Localization and 0.89 SODA_c on Dense Video Captioning, reported in their respective metrics rather than compared. In Dense Video Captioning, models generate accurate semantic descriptions yet fail to localize them in time, depressing the joint SODA_c metric; we separate the two components in Appendix D.

Spatio-Temporal Understanding. Single Object Tracking shows the widest separation between proprietary and open-weight models of any task. GPT-6 Astra reaches 81.20 Success AUC, level with the best specialist tracker at every horizon, and Gemini 3.6 Flash 75.99, against 64.66 for the strongest open-weight model, Cosmos3-Super. The spread within the open-weight group is itself large, from 64.66 down to 11.58. Scored on the same frames the models see, a static box that never updates reaches 24.64 Success AUC and a randomly placed box 6.55 (Appendix D.4). Three of the eighteen models score below the static box and a fourth is level with it. Qwen3.5-27B is the clearest case, second in the suite on 2D Object Localization at 85.46 yet below a box that never moves: single-frame spatial competence does not transfer to temporal persistence.

4.3 The Operational Deployment Gap (Δ)

To test whether the Infrastructure AI Gap is real, we compare model performance on VANTAGE-Bench against the same models' reported baselines on standard, consumer-focused datasets (Table 2). We include every model whose developer publishes reference scores under a matching metric: Qwen3-VL reports all four at 2B, 8B, and 32B. Frontier proprietary releases and newer open-weight families increasingly report only high-level multiple-choice cognitive benchmarks and no longer publish zero-shot baselines for granular spatial and temporal grounding, which is itself symptomatic of the evaluation gap we describe.

The deficit is localized, not uniform. The deltas run in both directions, and that is what makes the pattern informative. Event verification and temporal localization drop roughly 12 to 24 points at every scale we could match, the three Qwen3-VL checkpoints for which the developer published comparable numbers. Video question answering sits within 5.3 points, and pointing is flat at the two smaller scales (-0.7 and -0.5) before turning into a surplus of $+8.3$ at 32B. The same models therefore lose ground on some capabilities and gain it on others within one benchmark, which locates the shortfall in specific capabilities rather than in overall difficulty.

		Qwen3-VL		
Task	**Reference benchmark**	**2B**	**8B**	**32B**
VQA	VideoMME [1]	+2.0	−5.0	−5.3
Event Verification	MLVU [32]	−23.5	−18.6	−18.3
2D Spatial Pointing	BLINK [4]	−0.7	−0.5	+8.3
Temporal Localization	Charades-STA [11]	−19.3	−11.7	−14.4

Table 2: The Infrastructure AI Gap (Δ = VANTAGE − reference), for the four tasks where a developer-published reference exists under the same metric. Event verification is compared on accuracy to match the reference; Table 1 reports Macro F1 as that task's primary metric.

Where the shortfall sits. Three observations locate it. Models sharing one base diverge sharply by training focus (Section 4.4), so the pattern moves with what a model was trained on rather than with the suite as a whole. Proprietary access is no shortcut either: two proprietary models fall below four open-weight models on the overall pillar average (Table 1). Most directly, the split appears

within a single task. On dense video captioning, BERTScore-F1 spans 54.2 to 63.9 across models while temporal IoU-F1 spans 33.2 to 42.7 on the same items and the same annotations (Table 5): models describe the events correctly and misplace them in time. The failure is in temporal grounding specifically, not in video understanding broadly.

A prompt- and metric-matched control. Our temporal localization evaluation uses the prompt template the developer published for its own Charades-STA evaluation, so task, metric, and prompt are held constant. Qwen3-VL drops from 56.0 to 44.3 at 8B and from 61.2 to 46.8 at 32B.

A capability with no external baseline. The general-purpose models we evaluate claim tracking in their demonstrations, as the Qwen3-VL release does, but none of their releases reports a quantitative tracking benchmark, and to our knowledge no prior VLM tracking evaluation targets fixed-camera infrastructure viewpoints. Frontier models have therefore been deployed into physical environments without a published baseline for continuous spatio-temporal attention. VANTAGE-Bench supplies one, anchored against specialist trackers in Appendix D.4; our sequences are synthetic, and real fixed-camera tracking sequences are a planned extension (Section 5).

4.4 Returns to Scale and the Temporal Lag

A prevailing hypothesis is that grounding deficits will resolve through parameter scaling. The four open-weight scaling series let us test this directly (Table 7).

Returns to scale are task-specific, not pillar-specific. Scaling helps substantially on some tasks and not at all on others, and the pattern does not align with the four pillars. Temporal Localization improves by an average of 7.5 points from the smallest to the largest family member, so temporal grounding does respond to scale. Object Localization improves with scale in three of the four families; Cosmos-Reason2-32B is the exception, due to an output-format regression rather than a perception result. Qwen3.5 regresses on four of eight tasks despite tripling in size.

The persistent finding is an absolute temporal lag. GPT-6 Astra exceeds the next-best observed Temporal Localization score by 9.39 mIoU and the next-best Dense Video Captioning score by 0.89 SODA_c. We report these changes in their respective metrics without treating them as a calibrated comparison of capability progress. Its margin over the next model runs 3.0 to 5.2 on the other six tasks, except 2D Object Localization, where it trails the best open-weight model by 4.8. Temporal grounding on operational footage is therefore unsolved at present scale, though nothing in our evidence bounds what further scaling might achieve.

Physical-AI training data helps some tasks far more than others. Qwen3-VL, Cosmos-Reason2, and Cosmos3 share a base lineage and carry progressively more physical-AI training [43, 38], though the composition of that mixture is not public. Comparing each family's best score on each task, tracking gains 20.5 points across the progression and object localization 14.3, while temporal localization gains 5.1 and dense video captioning 1.8. Tracking's path is uneven: Cosmos-Reason2 falls below Qwen3-VL before Cosmos3 overtakes both. All models are evaluated zero-shot. Our working hypothesis is a training-mixture asymmetry: bounding-box supervision is abundant while densely timestamped video is scarce. Distinguishing a capability gap from an elicitation gap would require a targeted fine-tuning probe.

4.5 Two Tracking Failure Modes

Success AUC alone conflates two mechanically different failures. We therefore report a *freeze rate*: the fraction of predicted boxes within IoU ≥ 0.95 of the previous frame's prediction. At a 32-frame horizon Cosmos-Reason2-8B holds a 93% freeze rate at 18.0 AUC, having largely stopped updating its prediction, while Qwen3-VL-8B holds 49% at 15.7 AUC, continuing to update but drifting off target. The two scores are close; the failures are not. Specialist trackers sit at 16–31% and a static box at 99%, bracketing the metric at both ends. Appendix D.4 reports these systems at three horizons.

5 Discussion and Limitations

Coverage. Our claims are scoped to the configurations actually sampled: elevated fixed cameras at 8 to 20 feet and 30 to 60 degrees from vertical for the video tasks and 2D Object Localization, oblique aerial views for 2D Referring Expressions, and vehicle-mounted footage for 2D Spatial Pointing. Eye-level retail and low-angle industrial viewpoints are not covered; the constraint is data rights, since each deployment type requires separate licensing. Real-world footage comes primarily from two U.S. municipalities with supplementary licensed European footage, so coverage is weighted toward North America.

Annotation. Counts are uneven across the three regimes because the tasks differ in density: infrastructure scenes contain many simultaneous people and vehicles, so a single detection image carries 43.6 boxes on average. We therefore report the regimes separately rather than as one aggregate. Our acceptance rates come from sequential second-expert review, in which a reviewer corrects a prior annotation rather than labeling it independently; they measure how often expert work passes review, not agreement between independent annotators.

Construction caveats. For 2D Spatial Pointing the correct answer is fixed by human-labeled geometry and cannot be altered by the rewriting stage, but the rewritten text is not constrained to preserve the uniqueness of its spatial cue in dense scenes, and we do not verify that it does. The Single Object Tracking track is built *entirely* from synthetic warehouse sequences and contains no real footage, so its conclusions do not yet separate tracking failure from synthetic-domain mismatch; the specialist-tracker and horizon analyses in Appendix D.4 establish that the task is well posed but do not substitute for real-data validation. Real fixed-camera tracking sequences will be added in a future release.

Model coverage and reported uncertainty. Evaluation is strictly zero-shot; domain-specific fine-tuning is left to future work, and extreme weather and lighting are not systematically covered. Inference settings were tuned per model following each developer's published guidance where it exists; Appendix F.2 reports the configuration each model was finally evaluated under rather than the alternatives we tried. The intervals we report (Appendix G.1) cover item sampling only, not prompt or decoding sensitivity.

Planned extensions. We are extending every task with real-world footage across all three deployment domains, adding real fixed-camera tracking sequences, and planning a human performance baseline on a representative subset to calibrate task difficulty independently of annotation quality.

6 Conclusion

We introduced VANTAGE-Bench, an evaluation suite for VLMs in open-loop Infrastructure AI settings. Moving beyond the priors of internet video and single-format MCQ evaluation, we built an 8-task taxonomy spanning semantic, spatial, temporal, and spatio-temporal reasoning over 3,346 expert-annotated media assets, and evaluated 18 models zero-shot.

The resulting picture is uneven rather than uniformly poor. Open-weight models lead 2D Object Localization outright, the one task where the newest frontier model does not. The deficit against consumer-centric benchmarks concentrates in event verification and temporal grounding, while pointing is flat at smaller scales and turns into a surplus at 32B. Returns to scale are task-specific rather than pillar-specific. What persists is a lag on the two temporal tasks: GPT-6 Astra leads the field by 9.39 mIoU on temporal localization and 0.89 SODA_c on dense video captioning, reported in their respective metrics. On tracking, the separation from specialist trackers widens with horizon for every system except the newest. Astra leads seven of the eight tasks, which makes it the standing target for open-weight systems to close on; that it improves on GPT-5.6 Sol across all eight, by 0.9 to 12.9 points, shows the suite still registers frontier progress rather than having been saturated by it. We hope VANTAGE-Bench provides a useful framework for evaluating physical reasoning in operational settings, and that its open leaderboard makes progress on these specific deficits measurable.

Reproducibility Statement

The evaluation harness, including our metric scripts and the per-task prompts, is available for review at an anonymized repository: `https://github.com/anon-benchmark/VANTAGE-bench`. All evaluations are zero-shot and run through that single harness (Appendix F); decoding is greedy throughout, and Appendix F.2 reports the per-task frame-rate, frame-count and pixel budgets, within which sampling rate, resolution and prompt were tuned per model following each developer's published guidance. Formal task definitions and metric formulations are given in Appendix B. We release the input media, all prompts and questions, and the evaluation harness; ground-truth annotations are held out and scored by the leaderboard server, a protocol described in Appendix F.3 and chosen to keep the benchmark uncontaminated. Item-level bootstrap intervals for every model and task, together with a seed-variance study, are reported in Appendix G, so reported differences can be weighed against measurement error. The de-identification pipeline and consent provenance of the media are documented in Appendix E.

Ethics Statement

The automation of fixed-camera monitoring carries privacy implications. All raw assets underwent a de-identification pipeline combining automated obfuscation with human verification, documented in Appendix E. VANTAGE-Bench is released under a license prohibiting biometric identification and demographic profiling. We emphasize that the benchmark evaluates physical and causal reasoning rather than identity recognition, and we advocate for human-in-the-loop oversight in downstream operational deployments.

Use of Large Language Models

Language models were used both inside the benchmark pipeline and in preparing this paper.

Benchmark construction. One track uses a model in the loop: the Pointing Data Engine has a VLM paraphrase the question text for 2D Spatial Pointing, while the question's correct answer derives deterministically from human-labeled box geometry and cannot be altered by that stage (Section 3.3). No other track's questions, answers, or annotations were model-generated; all video and detection annotations were produced by trained human annotators under the protocol in Section 3. All evaluation is zero-shot, so no model-generated data was used for training.

Paper preparation. The authors used LLM assistance for drafting, editing, and condensing prose. Every reported number, claim, and citation was verified by the authors against the evaluation outputs and the cited sources.

Research content. The research question, benchmark design, task taxonomy, annotation protocol, experimental methodology, and interpretation of results are the authors' own. The paper makes no theoretical claims and contains no proofs. Analysis code was written by the authors with LLM assistance on routine implementation; all results were produced by the evaluation harness described in Appendix F.

REFERENCES

[1] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[2] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[3] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[4] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[5] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[6] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yuxin Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[7] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo Open Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[8] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[9] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, J.K. Aggarwal, Hyungtae Lee, Larry Davis, et al. A large-scale benchmark dataset for event recognition in surveillance video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[10] Zheng Tang, Shuo Wang, David C. Anastasiu, Ming-Ching Chang, Anuj Sharma, Quan Kong, Munkhjargal Gochoo, Jun-Wei Hsieh, Tomasz Kornuta, et al. The 10th AI City Challenge. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2026.

[11] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[12] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[13] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[14] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. LaSOT: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[15] Lianghua Huang, Xin Zhao, and Kaiqi Huang. GOT-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1562–1577, 2021.

[16] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. MOT16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016.

[17] Xiao Wang, Xiujun Shu, Zhipeng Zhang, Bo Jiang, Yaowei Wang, Yonghong Tian, and Feng Wu. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[18] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[19] Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[20] Han Wang, Yanjie Wang, Yongjie Ye, Yuxiang Nie, and Can Huang. Elysium: Exploring object-level perception in videos via MLLM. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[21] En Yu, Liang Zhao, Yana Wei, Jinrong Yang, Dongming Wu, Lingyu Kong, Haoran Wei, Tiancai Wang, Zheng Ge, Xiangyu Zhang, and Wenbing Tao. Merlin: Empowering multimodal LLMs with foresight minds. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[22] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. MMT-Bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask AGI. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

[23] Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. VideoChat-R1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025.

[24] Biao Wang, Wenwen Li, and Jiawei Ge. R1-Track: Direct application of MLLMs to visual object tracking via reinforcement learning. *arXiv preprint arXiv:2506.21980*, 2025.

[25] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations (ICLR)*, 2026.

[26] Cheng-Yen Yang, Hsiang-Wei Huang, Wenhao Chai, Zhongyu Jiang, and Jenq-Neng Hwang. SAMURAI: Motion-aware memory for training-free visual object tracking with SAM 2. *IEEE Transactions on Image Processing*, 35:970–982, 2026.

[27] Xin Chen, Ben Kang, Wanting Geng, Jiawen Zhu, Yi Liu, Dong Wang, and Huchuan Lu. SUTrack: Towards simple and unified single object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[28] Ben Kang, Xin Chen, Simiao Lai, Yang Liu, Yi Liu, and Dong Wang. Exploring enhanced contextual information for video-level object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[29] Zhichao Sun, Yepeng Liu, Zhiling Su, Huachao Zhu, Yuliang Gu, Yuda Zou, Zelong Liu, Gui-Song Xia, Bo Du, and Yongchao Xu. RefDrone: A challenging benchmark for referring expression comprehension in drone scenes. *arXiv preprint arXiv:2502.00392*, 2025.

[30] NVIDIA. PhysicalAI-SmartSpaces: A multi-camera 3D perception dataset for smart spaces. https://huggingface.co/datasets/nvidia/PhysicalAI-SmartSpaces, 2025.

[31] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. VLMEvalKit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM)*, pages 11198–11201, 2024.

[32] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. MLVU: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[33] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[34] Soichiro Fujita, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. SODA: Story oriented dense video captioning evaluation framework. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[35] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*, 2020.

[36] NVIDIA. NVIDIA Omniverse DRIVE Sim. https://developer.nvidia.com/drive/simulation, 2025.

[37] Qwen Team. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[38] NVIDIA. Cosmos3: Reasoning vision-language models for physical AI. https://huggingface.co/nvidia/Cosmos3-Super, 2026.

[39] Google DeepMind. Gemma 4 model card. https://huggingface.co/google/gemma-4-e2b-it, 2026.

[40] OpenAI. GPT-6 Astra. Model release, 2026.

[41] OpenAI. GPT-5.6 Sol. https://openai.com/index/gpt-5-6/, 2026.

[42] Google DeepMind. Gemini 3.1 Pro model card. https://deepmind.google/models/model-cards/gemini-3-1-pro/, 2026.

[43] NVIDIA. Cosmos-Reason2: An open reasoning vision-language model for physical AI and robotics. https://docs.nvidia.com/cosmos/latest/reason2/index.html, 2026.

[44] Qwen Team. Qwen3.5 model release. https://huggingface.co/Qwen/Qwen3.5-9B, 2026.

[45] Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations. *arXiv preprint arXiv:2411.00640*, 2024.

[46] Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, et al. TOMATO: Assessing visual temporal reasoning capabilities in multimodal foundation models. In *International Conference on Learning Representations (ICLR)*, 2025.

[47] Yuyou Zhang, Radu Corcodel, Chiori Hori, et al. SpinBench: Perspective and rotation as a lens on spatial reasoning in VLMs. In *International Conference on Learning Representations (ICLR)*, 2026.

[48] David Ma, Huaqing Yuan, Xingjian Wang, et al. ScaleLong: A multi-timescale benchmark for long video understanding. In *International Conference on Learning Representations (ICLR)*, 2026.

A Data Statistics and Benchmark Comparison

This appendix places VANTAGE-Bench alongside the benchmarks it is most often compared against and records how its annotations are distributed. Table 3 reports scale, modality, and pillar coverage for each.

Benchmark	Modality	# Media	# Annot.	Pillars	Annotation Source
VideoMME	Video	900	2,700	Semantic	Human
BLINK	Image	3,683	1,906	Spatial	Human, Existing
RefCOCO_avg	Image	3,982	30,969	Spatial	Human, Existing
ODinW13 [33]	Image	4,608	10,966	Spatial	Human, Existing
CharadesSTA	Video	1,334	3,720	Temporal	Human, PL
ActivityNet Cap.	Video	5,044	17,750	Temporal	Human
VANTAGE-Bench	**Image + Video**	**3,346**	**35,027**†	**All four**	Human, Simulated, PL, Existing

Table 3: Cross-benchmark comparison. All statistics reflect available evaluation test sets. *Pillars* is the reasoning taxonomy each benchmark covers. †Shown as a single total for comparability with the other rows; we report our annotations by regime elsewhere.

Each of the three annotation regimes listed in Section 3 is comparable to or larger than its established counterpart in Table 3: the video tasks carry more annotations than VideoMME (3,342 against 2,700), though over slightly fewer videos (854 against 900), and image grounding exceeds BLINK's test set (4,281 against 1,906). Tracking is the one smaller track. Detection dominates the aggregate because infrastructure scenes are dense, at 43.6 boxes per image against ODinW-13's 2.4, which reflects scene density rather than broad sampling.

B Formal Task Formulations and Evaluation Metrics

In this section, we provide the rigorous definitions for the mappings and evaluation protocols used in VANTAGE-Bench.

B.1 Notation and Definitions

Each task is stated in Table 4 as a function from inputs to a structured output, with its primary metric. We define the following input and output spaces to standardize the task formulations:

- **Inputs:** $\mathcal{I}$ (Image); $\mathcal{V}$ (Video); $\mathcal{Q}$ (Query); $B_{t=0}$ (Initial Box).
- **Outputs:** τ (Temporal segment $[t_{start}, t_{end}]$); C (Caption); a (Selection from candidates); $\{B\}$ (Set of coordinates).

B.2 Metric Clarifications

- **SODA$_c$** [34]**:** Evaluates narrative quality by matching predicted event sequences to ground truth based on temporal overlap and BERTScore-based [35] linguistic similarity.
- **Success AUC:** For SOT, the Success Plot counts frames where predicted IoU exceeds threshold σ, calculated by varying $\sigma \in [0, 1]$.
- **Macro F1:** Penalizes models that exploit majority-class bias in binary Event Verification.

C Pipeline Details and Data Adaptation

C.1 Safely Simulating Severe Anomalies

A fundamental challenge in Infrastructure AI is evaluating safety-critical anomalies (e.g., collisions, industrial accidents) that are inherently rare and heavily restricted by privacy regulations. To evaluate these high-stakes scenarios without ethical compromise, approximately 20% of our VQA and Temporal Localization splits utilize high-fidelity synthetic data generated using NVIDIA Omniverse

Pillar	Task	Mathematical Formulation	Primary Metric	Additional
Semantic	EV	$f(\mathcal{V}, \mathcal{Q}_{hyp}) \to \{0, 1\}$	Macro F1	Sens., Spec.
	VQA	$f(\mathcal{V}, \mathcal{Q}, \{A, B, C, D\}) \to a \in \{A, B, C, D\}$	Accuracy	—
Spatial	RefEx	$f(\mathcal{I}, \mathcal{Q}) \to [x, y, w, h]$	mIoU	—
	Pointing	$f(\mathcal{I}, \mathcal{Q}, \{A, B, C, D\}) \to a \in \{A, B, C, D\}$	Accuracy	—
	Obj Loc.	$f(\mathcal{I}, \mathcal{Q}_c) \to \{B_k\}_{k=1}^{M}$	F1@0.5	COCO mAP
Temporal	Temp Loc.	$f(\mathcal{V}, \mathcal{Q}) \to [t_{start}, t_{end}]$	mIoU	Recall@0.5
	DVC	$f(\mathcal{V}) \to \{(\tau_i, C_i)\}_{i=1}^{N}$	SODA_c	BERTScore-F1, IoU-F1
Spatio-Temp.	SOT	$f(\mathcal{V}, B_{t=0}) \to \{B_t\}_{t=1}^{T}$	Success AUC	Mean IoU

Table 4: Formal task definitions for VANTAGE-Bench. The primary metric is the one reported in Table 1 and on the public leaderboard.

DRIVE Sim [36]. This ensures VANTAGE-Bench covers severe long-tail physical anomalies that are otherwise absent from purely real-world datasets.

Benchmark Scope. Unlike prior datasets, which are typically restricted to a single modality or reasoning dimension, VANTAGE-Bench jointly evaluates semantic, spatial, temporal, and spatio-temporal reasoning across both image and video domains, and moves past MCQ-only evaluation to generative captioning and coordinate-based grounding. Appendix A compares its scale and coverage against representative benchmarks.

C.2 Data Adaptation for SOT and RefEx

2D Referring Expressions: We utilized RefDrone because its elevated perspective mimics infrastructure cameras. We ported prompts requiring dense disambiguation between identically colored vehicles based on relative spatial relationships.

Single Object Tracking: Adapted from the MTMC 2025 subset of PhysicalAI-SmartSpaces. We filtered 17 warehouse scenes featuring workers and robotic profiles (e.g., Nova Carter). Sequences include an initial "visual anchor" followed by a sequence of frames sampled at densities of 8, 16, and 32 frames.

D Extended Analysis: Deconstructing the Infrastructure AI Gap

Section 4 establishes the magnitude of the Infrastructure AI gap; the analyses below isolate the mechanisms behind it.

D.1 Semantic Illusion: Fluency vs. Grounding

We observe a systemic disconnect between linguistic fluency and physical grounding, which we term the **Semantic Illusion**.

Dense Video Captioning. Models generate semantically coherent descriptions (high BERTScore-F1) but fail to temporally localize events, resulting in low joint performance (SODA_c). This indicates that models capture *what* happens but fail to resolve *when* it occurs, leading to temporal smearing across events. Table 5 separates the two components. GPT-6 Astra achieves higher temporal IoU-F1 and SODA_c despite lower BERTScore-F1 than Gemini 3.1 Pro, consistent with improved temporal alignment contributing to its joint score.

Event Verification: Causal Blindness. The same disconnect appears in Event Verification, though the direction of the imbalance depends on the model rather than on scale alone (Table 6). Qwen3-VL-8B fails to detect true events while maintaining strong specificity, at 50.96 against 74.58, a

Model	Semantic Fluency (BERTScore-F1)	Discrete Loc. (IoU-F1)	Joint Metric ($SODA_c$)
GPT-6 Astra	61.39	**42.74**	**38.17**
Gemini 3.1 Pro	**63.86**	39.44	35.55
Cosmos-Reason2-8B	58.08	36.70	32.50
Qwen3-VL-8B	58.66	33.24	29.64
Cosmos-Reason2-32B	54.74	33.95	30.04
Cosmos3-Super	54.24	33.39	29.54

Table 5: Dense Video Captioning decomposed into its semantic and temporal components, with the joint metric for reference. Scaled 0–100.

bias toward conservative predictions under uncertainty rather than an inability to reason about the scene. Cosmos-Reason2-8B shows the opposite profile, at 71.15 sensitivity against 57.63 specificity, accepting events that did not occur. Macro F1 conceals both, which is why we report the two components separately. The asymmetry is a property of the model rather than of the split: on the same items GPT-6 Astra is balanced, at 88.46 sensitivity against 86.44 specificity, and it is the only model here above 86 on both. What separates it from Gemini 3.6 Flash is almost entirely specificity, 86.44 against 76.27, rather than sensitivity, 88.46 against 87.50.

Model	Macro F1	True Event Recall (Sensitivity)	True Negative Recall (Specificity)
GPT-6 Astra	**86.90**	**88.46**	**86.44**
Gemini 3.6 Flash	82.00	87.50	76.27
GPT-5.6 Sol	76.14	77.88	76.27
Cosmos-Reason2-32B	73.58	83.65	62.71
Cosmos3-Super	71.28	72.12	72.88
Gemini 3.1 Pro	68.57	71.15	67.80
Cosmos-Reason2-8B	64.09	71.15	57.63
Qwen3-VL-8B	59.39	50.96	74.58

Table 6: Event Verification causal breakdown ($n = 163$; 104 positive, 59 negative). Sensitivity and specificity are reported separately, since Macro F1 alone conceals the asymmetry.

D.2 Returns to Scale by Family

Table 7 gives the per-family scaling series summarized in Section 4.4.

	Open-weight scaling series				*Proprietary*		
Task	**Qwen3-VL** (2B→32B)	**Qwen3.5** (9B→27B)	**Cosmos-Reason2** (2B→32B)	**Cosmos3** (Edge→Super)	**GPT-5.6** Sol	**Gemini 3.6** Flash	**GPT-6** Astra
VQA	63.9→71.3 (+7.4)	71.6→68.0 (−3.6)	64.7→70.3 (+5.6)	62.7→69.5 (+6.8)	78.1	76.8	**83.3**
EV	44.7→60.0 (+15.3)	46.8→55.3 (+8.5)	55.3→73.6 (+18.3)	63.4→71.3 (+7.9)	76.1	82.0	**86.9**
Obj Loc.	47.4→72.7 (+25.3)	76.9→85.5 (+8.6)	73.6→4.4† (−69.2)	79.0→**87.0** (+8.0)	69.3	81.6	82.2
RefEx	65.1→72.4 (+7.3)	77.2→76.4 (−0.8)	56.3→50.0 (−6.3)	69.3→76.2 (+6.9)	69.4	75.8	**80.2**
Pointing	53.1→75.6 (+22.5)	71.7→75.3 (+3.6)	59.7→74.0 (+14.3)	67.8→72.9 (+5.1)	81.9	79.2	**87.0**
Temp Loc.	35.2→46.8 (+11.6)	42.5→36.9 (−5.6)	39.0→50.5 (+11.5)	39.5→51.9 (+12.4)	55.7	51.5	**65.1**
DVC	19.9→29.4 (+9.5)	27.9→27.0 (−0.9)	28.5→30.0 (+1.5)	16.6→29.5 (+12.9)	37.3	36.2	**38.2**
SOT	30.4→44.2 (+13.8)	23.5→24.4 (+0.9)	26.9→33.1 (+6.2)	24.7→64.7 (+40.0)	75.0	76.0	**81.2**

Table 7: Returns to scale, smallest to largest member of each open-weight family, with the change in parentheses, alongside the three newest proprietary models. Row-best in bold; markers as in Table 1.

D.3 Spatial vs. Spatio-Temporal Breakdown

We identify a key architectural limitation: *static spatial grounding does not imply temporal persistence*. Table 8 contrasts the two for seven representative models.

The dissociation is sharpest in the open-weight group. Qwen3.5-27B reaches 85.46 on 2D Object Localization and 76.35 on 2D Referring Expressions, among the strongest spatial results in the suite, yet scores 24.40 on tracking. Cosmos3-Super posts comparable spatial numbers and reaches 64.66. Spatial grounding is therefore not sufficient for temporal persistence, and the two are not acquired together. The proprietary rows are the contrast: GPT-6 Astra is strongest on both, at 86.97 pointing and 81.20 tracking, and the dissociation does not appear in that group at all.

Model	*Spatial*			*Spatio-Temporal*
	Pointing (Accuracy)	**RefEx** (mIoU)	**Object Loc.** (F1@0.5)	**SOT** (AUC)
GPT-6 Astra	**86.97**	**80.18**	82.20	**81.20**
GPT-5.6 Sol	81.90	69.44	69.30	75.02
Gemini 3.6 Flash	79.20	75.78	81.57	75.99
Cosmos3-Super	72.94	76.23	**86.97**	64.66
Qwen3.5-27B	75.32	76.35	85.46	24.40
Cosmos-Reason2-8B	68.60	70.26	83.88	37.69
Qwen3-VL-8B	68.56	73.29	59.93	33.12

Table 8: Single-frame spatial scores against Single Object Tracking, for seven representative models. Metrics scaled 0–100.

D.4 Tracking Anchors and Horizon Analysis

Every system in Table 9 is reported at all three horizons.

Protocol. All evaluations are performed on the same underlying object tracklets (200 sequences); the 8-frame, 16-frame, and 32-frame benchmarks differ only in the temporal sampling and clip length. The 8f and 16f datasets share an identical frame-level prefix (stride 15), with the 16f clips extending the 8f sequences by eight additional sampled frames. The 32f benchmark is independently resampled from the same tracklets using a stride of 14, resulting in a longer temporal window rather than a strict continuation of the 16f clips. At the default sampling rate (30 FPS), the three settings correspond to approximately 3.5 s (8f), 7.5 s (16f), and 14.5 s (32f) of video. Freeze rate is defined in Section 4.5.

Baselines. We evaluate four specialist trackers, a static-box floor, and a random floor on the same 200 tracklets and the exact frames the VLMs see, using the identical Success-AUC scorer. We did not add classical trackers (e.g., KCF, CSRT): on standard benchmarks they are bounded above by the modern trackers and below by these floors, so they would not change where the VLM scores sit.

Reading the anchors. The specialist trackers reach approximately 80 Success AUC and hold steady as the horizon grows, establishing the level a purpose-built system reaches on these annotations. GPT-6 Astra reaches 81.20 at 8 frames, level with the strongest specialist at 80.91, and holds that level to 32 frames at a loss of 0.55 points: it is the only VLM we evaluate that holds its accuracy across horizons. A single forward pass with no rolling memory therefore reaches the accuracy of stateful specialist trackers, so the protocol does not itself set the ceiling on this track. The preceding frontier generation does not (GPT-5.6 Sol -4.1, Gemini 3.6 Flash -9.7 from 8f to 32f). Open-weight behavior is heterogeneous. Cosmos3-Super remains well above the static floor at 32 frames (45.3 versus 13.8 AUC), although it drops 19.3 points as the horizon increases. Cosmos-Reason2-8B and Qwen3-VL-8B finish much closer to the static floor, at 18.0 and 15.7 versus 13.8. Among the VLMs, long-horizon degradation is therefore the rule everywhere except the newest frontier model, while convergence toward the static baseline is confined to the weaker open-weight models. The tier ordering—specialists and GPT-6 Astra together at the top, then the preceding frontier generation, the strongest open-weight model, then the 8B-class models—is unchanged across all three horizons.

Reading the freeze rate. Freeze rate must be interpreted jointly with AUC, since a nearly unchanged box can be correct when the target is stationary; freeze rate alone is not an error rate. The specialist trackers provide an empirical reference that itself rises with horizon: 16–18% at 8 frames against 27–31% at 16 and 32 frames. GPT-6 Astra sits above that reference at every horizon, at 23.75%,

Model	Success AUC ↑			Freeze rate			Δ AUC
	8f	16f	32f	8f	16f	32f	
Floors							
Random	6.55	6.27	6.12	0.00	0.00	0.00	−0.43
Static box	24.64	18.28	13.75	99.08	99.35	99.20	−10.89
Specialist trackers							
SAM 3	79.92	**82.25**	80.47	17.58	30.66	29.83	+0.55
SAMURAI	80.91	81.16	79.85	18.00	30.23	30.01	−1.06
SUTrack	79.12	81.30	78.50	16.33	28.43	27.77	−0.62
MCITrack	77.74	79.44	76.61	16.67	27.53	27.29	−1.13
Vision-language models							
GPT-6 Astra	**81.20**	81.84	**80.65**	23.75	37.94	38.02	−0.55
Gemini 3.6 Flash	75.99	70.77	66.25	23.75	36.46	36.66	−9.74
GPT-5.6 Sol	75.02	74.75	70.91	22.75	37.32	34.96	−4.11
Cosmos3-Super	64.66	61.32	45.32	32.08	48.74	58.99	−19.34
Cosmos-Reason2-8B	37.69	30.64	17.98	77.67	84.57	92.96	−19.71
Qwen3-VL-8B	33.12	23.27	15.67	38.42	47.20	48.50	−17.45

Table 9: Tracking anchors and horizon analysis. Success AUC and freeze rate at three horizons, scaled 0–100. The 8-frame column is the Success AUC reported on the public leaderboard.

37.94% and 38.02%, while remaining level with the best specialist on Success AUC: a higher freeze rate is not by itself a worse tracker. The static-box floor, which by construction never updates, sits at 99% throughout, so the metric is bracketed at both ends. Cosmos-Reason2-8B has a much higher freeze rate of 93% together with only 18.0 AUC at 32 frames, consistent with a failure to update its prediction. Qwen3-VL-8B has a lower freeze rate of 49% but only 15.7 AUC, consistent with continued box updates that drift off target. We therefore use freeze rate to distinguish these two failure modes, without interpreting the specialist rate as the true fraction of stationary targets.

The track covers 200 objects across 102 videos, in line with established tracking test sets (GOT-10k validation: 180 videos; LaSOT: 280). What separates VLMs from purpose-built trackers on this track is persistence over time rather than box regression: GPT-6 Astra stays level with the best specialist at every horizon, while for the preceding generation the gap widens as the horizon extends—Gemini 3.6 Flash falls 9.7 points from 8 to 32 frames and Cosmos3-Super 19.3.

D.5 Domain-Specific Vulnerabilities

Model performance varies significantly across operational domains, revealing sensitivity to environmental priors (Table 10).

Domain	VQA (Acc %)	DVC (SODA$_c$)
Warehouse	75.73	40.61
Smart Spaces	55.38	30.38
Transportation	71.51	8.35
Overall	69.46	29.54

Table 10: Cosmos3-Super by operational domain, holding model and annotation pipeline fixed.

Unstructured Intent. Semantic reasoning is weakest in Smart Spaces, at 55.38 VQA accuracy against 75.73 in Warehouse, which we attribute to ambiguous human behaviour and the absence of structured, task-driven interactions.

Egocentric Bias is Temporal, not Semantic. Transportation leaves VQA accuracy nearly intact at 71.51 but reduces dense video captioning to 8.35 from 40.61 in Warehouse. The model still recognises what is in the scene; it loses the ability to say when events begin and end. We attribute this to the mismatch between ego-centric training footage and fixed-camera evaluation perspectives, which is

most disruptive to the continuous-motion cues temporal grounding depends on. The domain effect therefore falls on the temporal pillar rather than the semantic one.

D.6 Pillar Scores

The main results table reports the eight task scores and the *Overall* column, but not the intermediate pillar scores from which *Overall* is computed. Table 11 supplies them. Each pillar score is the unweighted mean of its constituent task scores, and *Overall* is the unweighted mean of the four pillar scores, so a pillar containing one task carries the same weight as a pillar containing three. This equal weighting is deliberate: it prevents the spatial pillar, which has the most tasks, from dominating the aggregate. It does mean the Spatio-Temporal pillar is identical to the Single Object Tracking column of Table 1, since tracking is its only task, and that tracking therefore carries a quarter of the *Overall* score. Readers who prefer a task-weighted aggregate can compute one from Table 1; it would raise the relative weight of the spatial pillar, where open-weight models are strongest.

Model	**Params**	**Semantic** (2 tasks)	**Spatial** (3 tasks)	**Temporal** (2 tasks)	**Sp.-Temp.** (1 task)	**Overall**
Open-weight models						
Qwen3-VL	2B	54.29	55.22	27.55	30.38	41.86
Qwen3-VL	8B	62.91	67.26	36.98	33.12	50.07
Qwen3-VL	32B	65.67	73.56	38.10	44.19	55.38
Qwen3.5	9B	59.20	75.26	35.20	23.52	48.30
Qwen3.5	27B	61.65	79.04	31.93	24.40	49.25
Cosmos-Reason2	2B	59.98	63.20	33.73	26.85	45.94
Cosmos-Reason2	8B	66.02	74.25	39.90	37.69	54.46
Cosmos-Reason2	32B	71.94	42.81†	40.25	33.12	47.03†
Cosmos3-Edge	4B	63.02	72.02	28.07	24.65	46.94
Cosmos3-Nano	16B	68.91	74.84	39.73	59.21	60.67
Cosmos3-Super	64B	70.37	78.71	40.72	64.66	63.62
Gemma-4-E2B	2B	36.92	27.55	9.08	11.58	21.28
Proprietary models						
Gemini 3.1 Flash-Lite	—	65.91	62.71	35.13	46.68	52.61
Gemini 3.1 Pro	—	70.01	72.14	40.62	67.88	62.66
Gemini 3.5 Flash-Lite	—	68.01	61.09	28.14	55.77	53.25
Gemini 3.6 Flash	—	79.41	78.85	43.83	75.99	69.52
GPT-5.6 Sol	—	77.11	73.55	46.50	75.02	68.04
GPT-6 Astra	—	**85.08**	**83.12**	**51.63**	**81.20**	**75.26**

Table 11: Pillar scores and their aggregate. Each pillar is the unweighted mean of its tasks; *Overall* is the unweighted mean of the four pillars. The Spatio-Temporal pillar consists of Single Object Tracking alone. All values scaled 0–100; bold denotes the best score per column. Markers as in Table 1.

D.7 Where the Frontier Advantage Actually Lies

The frontier advantage is largest on tracking and smallest on spatial grounding, and this section sets out the evidence for that.

There is no across-the-board proprietary advantage. Open-weight models lead 2D Object Localization outright, at 86.97 against 82.20 for the strongest proprietary model—a gap whose item-level intervals do not overlap (Table 13)—and two proprietary models fall below four open-weight models on the overall pillar average (Table 1). Proprietary access alone does not confer an edge.

The advantage belongs to the newest releases. Proprietary models lead the best open-weight model on seven of eight tasks, by margins that are large but not uniform: 3.0 on referring expressions, 5.7 on dense video captioning, 11.4 to 13.3 on pointing, VQA and event verification, 13.2 on temporal localization, and 16.5 on tracking. Object localization is the lone exception, where Cosmos3-Super leads by 4.8 points.

The distinctive frontier capability is temporal persistence. The tracking gap is where the frontier separates most sharply, at 16.5 points, and the horizon analysis in Appendix D.4 sharpens the picture: GPT-6 Astra is essentially flat across horizons, losing 0.6 points from 8 to 32 frames, the preceding frontier generation loses 4 to 10, and weaker open-weight models converge toward the static-box floor.

E De-identification Pipeline

What counts as PII. We treat as personally identifiable any information that can identify or locate a person: names, recognizable faces, license plates, addresses, and GPS or location data.

Detection and obfuscation. Faces, license plates, and video locations were obfuscated using a proprietary detector. Detector output was then verified and corrected by human annotators, so that the released media reflect a human-in-the-loop pass rather than automated obfuscation alone.

Licensing safeguards. Beyond obfuscation, the release license prohibits biometric identification and demographic profiling, restricting the released data to evaluation and benchmarking use.

Consent. 70% of footage was provided by vendors who obtained explicit informed consent from recorded individuals for research and redistribution. A smaller portion was captured in limited spaces with posted notice of camera operation, where individuals incidentally captured consented to fixed-camera monitoring as a condition of site access.

F Reproducibility and Experimental Configuration

F.1 Compute Resources and Frameworks

Evaluations for open-weight models (Cosmos-Reason2, Qwen3-VL, Qwen3.5, Cosmos3, Gemma-4) were conducted on an academic HPC cluster, using 7 nodes of 8× NVIDIA H100 (80GB) GPUs, as well as additional cloud compute instances. Proprietary models (GPT-6 Astra, GPT-5.6 Sol, and the Gemini family) were accessed via public APIs. We build our evaluation harness on top of VLMEvalKit [31], extending it with custom metric scripts for automated SODA_c computation, temporal IoU evaluation, and spatial trajectory parsing.

F.2 Task-Specific Inference Configurations

All evaluations used greedy decoding at temperature 0. Sampling rate, resolution, and prompt were tuned per model following each developer's published guidance where it exists. Chain-of-thought was disabled for open-weight models (`enable_thinking = False`); the proprietary APIs expose no equivalent control. The sampling and pixel budgets used to manage visual granularity and token limits are given in Table 12. FPS ranges reflect model-specific sampling rates.

Pillar	Task	FPS	Max Frames	Max Total Pixels
Semantic	EV / VQA	1.0 – 4.0	256	16.7×10^6
Spatial	Pointing / RefEx	—	1	1.5×10^6
Temporal	Temporal Localization	2.0 – 10.0	256	16.7×10^6
Temporal	DVC	1.0 – 4.0	128	8.4×10^6
Spatio-Temp.	SOT	—	8 / 16 / 32	0.9×10^6
Spatial	2D Object Localization	—	1	1.5×10^6

Table 12: VANTAGE-Bench Task Inference Parameters.

F.3 Code, Data Access, and Release Strategy

The dataset, evaluation harness, and public leaderboard are publicly released. The harness is available for review at an anonymized repository, `https://github.com/anon-benchmark/`

VANTAGE-bench; the dataset and leaderboard URLs are withheld for double-blind review and will be included in the camera-ready version. To preserve long-term integrity against training-data contamination, we adopt a held-out evaluation protocol. We publicly release (i) all input media, (ii) all prompts and questions, and (iii) the evaluation harness. Ground-truth annotations are held-out to be used by the official leaderboard server. The leaderboard is live and open for submissions, which are uploaded as JSON in LLaVA format.

G Reported Uncertainty

We report two kinds of variability, which answer different questions: how precisely a score is estimated from the items in the test set, and how reproducible it is across runs.

G.1 Item-Level Confidence Intervals

Because all evaluation uses greedy decoding at temperature 0, run-to-run variance is small (Section G.2, below); the dominant source of uncertainty is which items happen to be in the test set. We therefore report 95% percentile bootstrap confidence intervals over test items, resampling items with replacement 10,000 times per model–task cell and recomputing the metric on each resample. For Single Object Tracking the resampling unit is the object tracklet ($n = 200$) rather than the individual frame, since frames within a tracklet are strongly correlated.

The half-width of each interval is given in Table 13. Interval width is governed by the number of scored items, so it varies far more across tasks than across models. Referring expressions and object localization are the tightest (median ± 0.99 and ± 1.09 respectively), while event verification is the widest by a wide margin (median ± 7.35), reflecting its smaller item pool and binary scoring. The practical consequence is that GPT-6 Astra's 3.02-point lead on 2D Referring Expressions is resolved by the intervals while its larger 4.90-point lead on Event Verification is not, so the comparisons we draw in Section 4 should be read against very different resolutions: a two-point difference is meaningful on referring expressions and is not meaningful on event verification.

		Semantic		*Spatial*			*Temporal*		*Sp.-Temp.*
Model	**Params**	**VQA**	**EV**	**Point.**	**RefEx**	**Obj Loc.**	**Temp Loc.**	**DVC**	**SOT**
Open-weight models									
Qwen3-VL	2B	63.85 ±2.76	44.73 ±7.56	53.13 ±3.08	65.12 ±1.10	47.42 ±2.68	35.21 ±2.23	19.90 ±1.77	30.38 ±2.83
Qwen3-VL	8B	66.44 ±2.68	59.39 ±7.64	68.56 ±2.94	73.29 ±1.01	59.93 ±3.11	44.32 ±2.49	29.64 ±2.26	33.12 ±2.90
Qwen3-VL	32B	71.30 ±2.55	60.04 ±7.66	75.62 ±2.64	72.39 ±1.00	72.67 ±2.51	46.80 ±2.40	29.40 ±2.87	44.19 ±2.98
Qwen3.5	9B	71.63 ±2.51	46.78 ±7.61	71.74 ±2.79	77.16 ±0.90	76.87 ±1.81	42.53 ±2.49	27.88 ±2.87	23.52 ±2.47
Qwen3.5	27B	67.95 ±2.68	55.34 ±7.73	75.32 ±2.64	76.35 ±0.89	85.46 ±0.64	36.88 ±2.22	26.97 ±2.53	24.40 ±2.35
Cosmos-Reason2	2B	64.69 ±2.72	55.27 ±7.97	59.70 ±3.08	56.33 ±1.24	73.56 ±1.62	38.95 ±2.53	28.52 ±3.12	26.85 ±2.71
Cosmos-Reason2	8B	67.95 ±2.72	64.09 ±7.70	68.60 ±2.89	70.26 ±1.02	83.88 ±0.89	47.30 ±2.53	32.50 ±3.50	37.69 ±3.12
Cosmos-Reason2	32B	70.29 ±2.59	73.58 ±7.00	74.03 ±2.74	50.04 ±1.05	4.35† ±0.91	50.47 ±2.65	30.04 ±3.18	33.12 ±3.02
Cosmos3-Edge	4B	62.68 ±2.76	63.36 ±7.51	67.76 ±2.84	69.27 ±1.09	79.02 ±0.70	39.51 ±2.57	16.64 ±2.85	24.65 ±2.81
Cosmos3-Nano	16B	68.95 ±2.68	68.88 ±7.16	74.83 ±2.69	75.57 ±0.99	74.11 ±1.74	48.04 ±2.61	31.42 ±3.81	59.21 ±3.75
Cosmos3-Super	64B	69.46 ±2.59	71.28 ±7.12	72.94 ±2.74	76.23 ±0.98	86.97 ±0.70	51.90 ±2.57	29.54 ±3.31	64.66 ±3.67
Gemma-4-E2B	2B	46.19 ±2.85	27.65 ±4.52	31.74 ±2.94	5.52 ±0.40	45.40 ±0.74	9.02 ±1.10	9.14 ±1.83	11.58 ±1.31
Proprietary models									
Gemini 3.1 Flash-Lite	—	68.03 ±2.72	63.78 ±7.47	63.68 ±2.99	52.70 ±1.41	71.74 ±1.34	37.72 ±2.36	32.54 ±2.46	46.68 ±3.89
Gemini 3.1 Pro	—	71.46 ±2.55	68.57 ±7.23	72.04 ±2.79	67.17 ±1.27	77.21 ±1.73	45.69 ±2.36	35.55 ±2.82	67.88 ±3.98
Gemini 3.5 Flash-Lite	—	65.61 ±2.68	70.41 ±7.06	43.08 ±3.03	74.46 ±0.97	65.73 ±1.27	24.00 ±2.67	32.29 ±2.39	55.77 ±4.56
Gemini 3.6 Flash	—	76.82 ±2.38	82.00 ±6.26	79.20 ±2.54	75.78 ±0.87	81.57 ±0.61	51.50 ±2.67	36.15 ±3.16	75.99 ±3.44
GPT-5.6 Sol	—	78.08 ±2.38	76.14 ±6.72	81.90 ±2.39	69.44 ±0.91	69.30 ±0.72	55.71 ±2.65	37.28 ±2.88	75.02 ±3.34
GPT-6 Astra	—	83.26 ±2.09	86.90 ±5.40	86.97 ±2.09	80.18 ±0.68	82.20 ±0.74	65.10 ±2.46	38.17 ±2.74	81.20 ±2.78

Table 13: Scores with 95% item-level bootstrap confidence-interval half-widths (±), for every cell of Table 1, in points on the 0–100 scale (10,000 resamples; tracking resampled at the tracklet level) [45]. Scored items per task: VQA 1195; EV 163; Point. 1005; RefEx 3276; Obj Loc. 628; Temp Loc. 203; DVC 104; SOT 200.

G.2 Seed Variance

To confirm that reported differences are not seed-driven, we ran Qwen3-VL-8B three times with different random seeds (Table 14). Because decoding is greedy at temperature 0, this variance is expected to be small, and it is. All five tasks measured show a standard deviation below 0.5 points,

with 2D Spatial Pointing and Single Object Tracking the largest at 0.42 and 0.40. This establishes that individual cells are reproducible across runs. It is not a basis for judging whether two models differ: seed variance holds the item set fixed and therefore omits the dominant source of uncertainty in benchmark scores, which is item sampling [45]. For model comparisons, the item-level intervals above are the relevant quantity.

Pillar	**Task (metric)**	**Std. dev. across 3 seeds**
Semantic	Event Verification (Macro F1)	0.31
Semantic	Video Question Answering (Accuracy)	0.09
Spatial	2D Spatial Pointing (Accuracy)	0.42
Temporal	Dense Video Captioning ($SODA_c$)	0.32
Spatio-Temp.	Single Object Tracking (AUC)	0.40

Table 14: Seed variance on VANTAGE-Bench (Qwen3-VL-8B), standard deviation across three random seeds, scaled 0–100. This reflects decoding randomness only, not item sampling.