

The 10th AI City Challenge

Zheng Tang¹, Shuo Wang¹, David C. Anastasiu²
 Ming-Ching Chang³, Anuj Sharma⁴, Quan Kong⁵
 Munkhjargal Gochoo⁶, Jun-Wei Hsieh⁷, Tomasz Kornuta¹
 Zhedong Zheng⁸, Renran Tian⁹, Judah Goldfeder¹⁰
 Fulgencio Navarro¹¹, Yuxing Wang¹, Yizhou Wang¹
 Sameer Satish Pusegaonkar¹, Anqi Li¹, Nalin Dadhich¹
 Ridham Kachhadiya², Dhanishtha Patil², Haoquan Liang¹
 Jiajun Li¹, Han Zhang¹, Yilin Zhao¹
 Zaid Pervaiz Bhat¹, Shuyu Yang¹², Ashutosh Kumar⁵
 Rong Wang⁵, Rafael Martin Nieto¹¹, Peter Christiansen¹¹
 Ahmed Abduljawad⁶, Mohanrasu Shanmugam⁶, Nadeem Shaik⁶
 Sujit Biswas¹, Xunlei Wu¹, Vidya Murali¹
 , and Rama Chellappa¹³

¹ NVIDIA ² Santa Clara University
³ University at Albany, SUNY ⁴ Iowa State University
⁵ Woven by Toyota ⁶ United Arab Emirates University
⁷ National Yang Ming Chiao Tung University ⁸ University of Macau
⁹ North Carolina State University ¹⁰ Columbia University ¹¹ Milestone Systems
¹² Xi'an Jiaotong University ¹³ Johns Hopkins University

Abstract. The 10th AI City Challenge, held with ECCV 2026, marks a decade of community benchmarking for intelligent transportation, smart cities, and physical AI. Since its 2017 start with vehicle detection, classification, and tracking, the challenge has grown into a broad benchmark suite for multi-camera perception, multimodal reasoning, synthetic-to-real learning, generative forecasting, and privacy-preserving evaluation. The 2026 edition continued this growth with 325 registered teams, up from 245 in 2025, and participation from 26 countries and regions, up from 15. Its six primary tracks cover multi-camera 3D perception, transportation safety captioning and VQA, traffic anomaly reasoning, text-based person anomaly search, generative traffic video forecasting, and cross-city object detection. Track 3 further includes two out-of-domain leaderboards, submitted as Tracks 7 and 8, for fisheye traffic-violation understanding and pedestrian situated-intent VQA. This paper summarizes the challenge setup, datasets, evaluation protocols, leaderboard results, and workshop papers. Across tracks, successful systems combine foundation models with geometric grounding, retrieval or reranking, synthetic-data design, domain adaptation, and controlled inference.

Keywords: AI City Challenge · Synthetic-to-real transfer · Video understanding · Intelligent transportation · Physical AI

1 Introduction

The AI City Challenge is a recurring benchmark venue for computer vision systems deployed in urban, transportation, and physical AI settings. Its 10th edition, hosted as an ECCV 2026 workshop, marks a decade of growth from the 2017 vehicle detection, classification, and tracking tasks to a wider testbed for multi-camera 3D perception, multimodal reasoning, synthetic-to-real transfer, generative prediction, and privacy-preserving evaluation.

This evolution parallels broader progress in AI. Early tasks reflected deep learning for detection, tracking, and re-identification; recent tasks reflect VQA, foundation models, generative video models, and reasoning systems that must explain events rather than only detect them. The 2026 cross-city object detection track also highlights a persistent limitation of modern AI systems: strong in-domain performance does not guarantee robustness under geographic, camera, and scene-domain shift.

Participation also reached a new high. The 2026 challenge registered 325 teams, compared with 245 in 2025, a growth of about 33%, and the participant pool represented 26 countries and regions, compared with 15 in the previous year. The breadth of participation is visible across the leaderboards. On the general and public leaderboards, respectively, Track 1 had 29 and 15 teams, Track 2 had 44 and 22, Track 3 had 76 and 27, Track 4 had 47 and 28, Track 5 had 18 and 12, and Track 6 had 39 and 29. The Track 3 out-of-domain leaderboards also drew participation, with 15 general and 8 public teams for FETV fisheye traffic-violation understanding and 15 general and 7 public teams for PSI-VQA pedestrian situated-intent VQA.

This edition therefore introduces a broader set of tasks than a single leaderboard can capture: multi-camera 3D perception, safety-oriented captioning and VQA, anomalous-event reasoning with two OOD leaderboards, text-based person anomaly search, text-conditioned future-frame generation, and cross-city object detection through Milestone Project Hafnia. Together, these tracks form an urban intelligence benchmark suite rather than isolated recognition tasks.

This summary paper follows the structure of recent AI City Challenge overview papers [40,54,64]: it first presents the challenge setup, then describes the datasets, evaluation protocols, and leaderboard system, and finally summarizes the track results and the main technical trends in workshop papers. Dataset and benchmark papers associated with the 2026 challenge are discussed in the dataset section, while the results section focuses on teams with identifiable evaluation-system entries.

2 Challenge Setup

The 2026 AI City Challenge was organized around six primary challenge tracks and eight evaluation-server leaderboards, because the main Track 3 task includes two optional out-of-domain evaluations submitted as Tracks 7 and 8 on the evaluation system. Teams registered, requested submission privileges, and submitted

predictions to public or general leaderboards [2]. Public leaderboard entries were intended for award-eligible submissions, while the general leaderboards recorded broader participation.

Track 1: Multi-Camera 3D Perception. Participants tracked people, robots, forklifts, pallet trucks, and humanoids across synchronized warehouse cameras. The task required per-frame 3D boxes, class labels, scene-level identities, and world-coordinate localization. Depth was available for training and validation only; hidden real-world testing required RGB-only inference.

Track 2: Transportation Safety Understanding and Captioning. Participants used synthetic Digital Twin WTS data to caption and answer questions about real WTS videos. Each event was divided into behaviorally meaningful phases, and systems generated pedestrian and vehicle captions while also answering multiple-choice questions about position, direction, attention, attributes, and context.

Track 3: Anomalous Events in Transportation. Participants built one unified system that detects, reasons about, and explains anomalous transportation events through binary and multiple-choice questions, open-ended explanation, causal linkage, scene description, temporal description, and summarization. Tracks 7 and 8 extended the same reasoning theme to fisheye traffic violations and pedestrian intent.

Track 4: Text-Based Person Anomaly Search. Participants retrieved real pedestrian images from natural-language descriptions that mention both appearance and behavior. The task is a synthetic-to-real retrieval problem: synthetic image-text pairs provide training signal, while the hidden real-world query-gallery test set measures whether models generalize to realistic abnormal and routine behaviors.

Track 5: Generative Traffic Video Forecasting. Participants generated future frames conditioned on recent history frames and textual descriptions of target future behavior. The task stresses temporal consistency, visual fidelity, and whether a generated traffic sequence remains semantically aligned with safety-critical pedestrian and vehicle behavior.

Track 6: Cross-City Object Detection. Participants trained object detectors through the Hafnia Training-as-a-Service platform, which exposes managed training and benchmarking workflows without allowing direct extraction of the full real-world traffic corpus. The hidden benchmark mixes source- and target-city samples to measure geographic and visual domain shift.

Across the tracks, the challenge combined classical perception metrics such as HOTA and mAP with multimodal language, retrieval, generation, and reasoning metrics. This design encouraged systems that combine foundation models with domain-specific constraints, rather than relying on a single model family across all tasks.



Fig. 1: Track 1 sample: synchronized multi-camera warehouse views and a top-down layout for RGB-only multi-camera 3D perception.

3 Datasets and Tracks

The 2026 challenge uses six primary tracks and two out-of-domain leaderboards on the shared evaluation system. The primary tracks are described on the challenge web site [3–8], while Tracks 7 and 8 extend Track 3 to fisheye traffic-violation understanding and pedestrian situated-intent VQA. The following subsections summarize the data settings, hidden-test assumptions, and evaluation focus for each benchmark.

3.1 Track 1: Physical AI Smart Spaces

Track 1 extends the PhysicalAI-SmartSpaces benchmark to multi-camera 3D perception in warehouse environments [3, 43]. The training and validation data were generated from a large set of simulated warehouse scenes with synchronized RGB and depth videos, calibrated camera poses, 2D and 3D annotations, top-down maps, and multi-object identities. The hidden test set used real-world video in layouts that differ from the synthetic training scenes, making the task a strict Sim2Real benchmark rather than a closed-world tracking exercise.

The central difficulty is the coupling between geometry and association. A system must detect objects in each camera, lift observations into a shared 3D coordinate frame, and maintain identities despite occlusion, similar instances, changing viewpoints, and cluttered warehouse geometry; Fig. 1 shows the synchronized views and top-down layout. The RGB-only test restriction prevents direct test-time depth use, so successful approaches rely on learned monocular cues, calibration-aware projection, scene priors, and global association.

3.2 Track 2: Digital Twin WTS

Track 2 uses the Digital Twin WTS setting for synthetic-to-real traffic safety understanding [4, 39, 48]. The training and validation data provide synthetic multi-view traffic-safety scenes, annotations, segment-level descriptions, and VQA labels, with representative synthetic and real views shown in Fig. 2. The hidden test set uses real WTS videos, so the benchmark measures whether methods can transfer from controlled synthetic scenes to real camera footage while preserving fine-grained safety semantics.

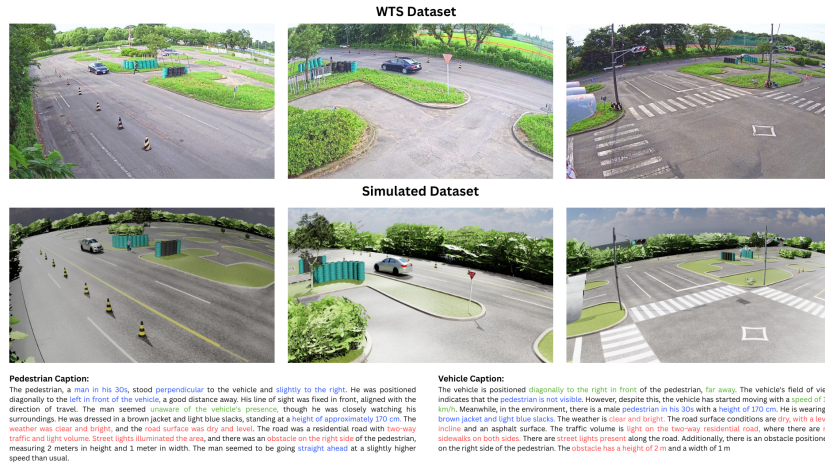


Fig. 2: Track 2 sample: synthetic Digital Twin WTS scenes paired with real WTS-style views and structured pedestrian/vehicle descriptions.

The task combines two complementary outputs. For captioning, systems describe pedestrian and vehicle behavior across event phases, including movement direction, relative position, gaze, visibility, attributes, weather, road geometry, and traffic context. For VQA, systems answer structured questions that test whether the model has grounded those descriptions in the visual evidence. The track therefore rewards representations that separate stable scene layout from variable appearance and behavior, rather than merely matching surface text patterns.

3.3 Track 3: Anomalous Events in Transportation

Track 3 introduces TAR (Traffic Anomaly Reasoning) and TAR-Bench for traffic anomaly reasoning [5, 44, 67]. The TAR training dataset contains 44,040 annotations over 3,670 transportation videos and covers 10 task types, including binary and multiple-choice reasoning, open-ended QA, captioning, causal explanation, scene description, temporal ordering, and summarization. TAR-Bench provides 960 human-curated annotations for 80 held-out clips. Unlike earlier anomaly-detection tasks, this track emphasizes explanation and evidence grounding: a strong model must identify what happened, infer why it happened, and describe the evidence coherently.

Two out-of-domain datasets extend Track 3. Track 7 evaluates FETV traffic-violation understanding from fisheye cameras [1], where geometry changes object scale, direction cues, and perspective. Track 8 evaluates PSI-VQA, a pedestrian situated-intent VQA benchmark for ambiguous pedestrian behavior around automated driving [63]. Fig. 3 illustrates the evidence, question, and explanation structure of the main TAR setting, and the evaluation system reports all three reasoning settings separately as Tracks 3, 7, and 8.



Fig. 3: Track 3 sample: traffic anomaly reasoning combines video evidence, task-specific questions, event descriptions, and causal explanations.

3.4 Track 4: Pedestrian Anomaly Behavior

Track 4 uses the Pedestrian Anomaly Behavior (PAB) benchmark for text-based person anomaly search [6]. The training set is synthetic and provides image-text pairs that describe person appearance, scene context, and action. The real test set contains query descriptions and gallery images. Participants rank gallery images for each query, and the final leaderboard uses retrieval quality on hidden real-world data.

The benchmark is difficult because abnormal behavior is often defined by relationships between action, body pose, scene context, and a natural-language query. Appearance-only retrieval is insufficient: examples such as falling, lying, being hit, or unsafe crossing require action grounding, as the synthetic/real and hard-negative examples in Fig. 4 illustrate. Successful methods therefore use text-image contrastive training, action-aware alignment, hard-negative mining, reranking, and query decomposition.

3.5 Track 5: Traffic Video Forecasting

Track 5 asks teams to generate future traffic video frames from a short history window and a textual description of expected future behavior [7]. The task builds on WTS-style traffic scenes but changes the output from analysis to generation, as shown by the history and target-frame example in Fig. 5. Models must synthesize plausible motion, preserve identity and background consistency, and reflect the target pedestrian/vehicle behavior.

This task exposes a different failure mode from captioning and retrieval: a generated video may be sharp but semantically wrong, or text-aligned while losing scene consistency and temporal smoothness. The leaderboard therefore combines low-level fidelity, perceptual, semantic, and video-distribution metrics. Accepted methods typically use diffusion or world-model priors, frame-history conditioning, text-guided planning, and post-generation selection.



Fig. 4: Track 4 sample: synthetic PAB training examples and real-gallery retrieval, including hard negatives that share appearance or scene context with the query.

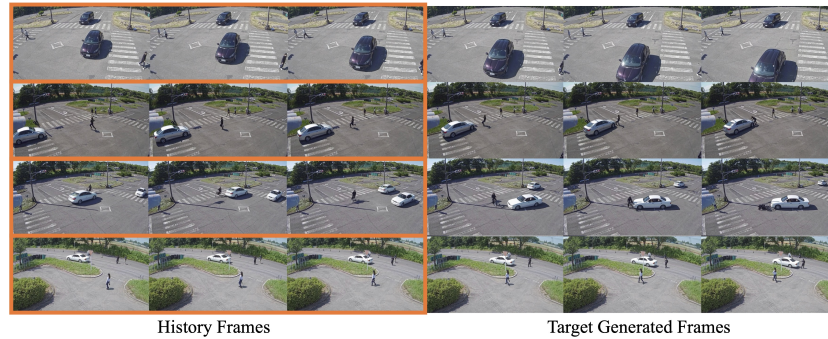


Fig. 5: Track 5 sample: future-frame generation from history frames and target behavior descriptions in traffic scenes.

3.6 Track 6: Hafnia Cross-City Detection

Track 6 uses Hafnia Training-as-a-Service platform [37] for cross-city object detection [8, 36]. The data consist of real traffic-camera images with object annotations across multiple categories. Unlike tracks where participants directly download the full corpus, this track uses a managed platform for privacy-preserving training, validation, and hidden benchmarking on controlled splits.

The track targets a common deployment problem: detectors trained in one city or camera network often lose accuracy when moved to another geography, camera height, lens, weather condition, or traffic pattern, as illustrated in Fig. 6. The hidden benchmark includes source- and target-city samples, rewarding detector design together with class-aware augmentation, resolution management, domain-shift validation, and confidence calibration.



Fig. 6: Track 6 sample: privacy-preserved traffic-camera images from Milestone Systems (Project Hafnia), with annotated source-city frames on top and hidden target-city frames below.

4 Evaluation Protocols

The evaluation system exposed public and general leaderboards for each track [2]. During the challenge, scores were computed on a subset of hidden test data and only limited ranking information was shown. After the deadline, final scores were recomputed on full hidden test sets and team names were made public for the workshop summary. Public/general ranks in the results tables use the form “1/21 (1/58)”, meaning rank 1 among 21 public teams and rank 1 among 58 general teams.

4.1 Track 1 Evaluation

Track 1 used 3D HOTA [35], which jointly evaluates detection accuracy, association accuracy, and localization quality over 3D tracks. Submissions provided class labels, frame indices, scene-level identities, and 3D boxes or equivalent localization fields in the warehouse coordinate frame. Scores were averaged across classes and scenes. Because the hidden test set was RGB-only, participants could not use test-time depth; online submissions received an additional ranking bonus when they used only current and past frames.

4.2 Track 2 Evaluation

Track 2 combined captioning and VQA. Caption quality used BLEU-4 [46], METEOR [10], ROUGE-L [32], and CIDEr [60]; VQA used answer accuracy over structured questions [9]. The final score averaged caption and VQA components, making the hidden real WTS test set a synthetic-to-real transfer benchmark for semantic grounding.

4.3 Track 3 Evaluation

Track 3 evaluated multi-task traffic anomaly reasoning. The official in-domain TAR-Bench mean was computed from nine scored task types, excluding temporal

localization. Closed-form tasks used accuracy-style scoring, while text-generation tasks used BLEU, METEOR, ROUGE, and CIDEr [10, 32, 46, 60]. The design rewards models that organize evidence over time and generate semantically correct, visually grounded answers.

4.4 Track 4 Evaluation

Track 4 used mean Average Precision (mAP), widely used in detection and retrieval benchmarks [16, 33], for text-to-image retrieval. For each text query, participants returned a ranked gallery list; the score rewards placing correct matches early while suppressing hard negatives. Because training images are synthetic and the test gallery is real, the metric measures both retrieval quality and Sim2Real generalization.

4.5 Track 5 Evaluation

Track 5 combined visual, perceptual, and semantic video metrics: PSNR and SSIM [65] for pixel fidelity, LPIPS [68] and CLIP-S [18] for perceptual and text-image alignment, and FID [19] and FVD [59] for distributional realism. The final score rewards realistic future frames that also follow the target behavior described in text.

4.6 Track 6 Evaluation

Track 6 used object detection mAP [16, 33] on hidden Hafnia benchmark images. The platform evaluated models on source-city and target-city samples, so the score reflected both in-domain detector quality and robustness to geographic shift. The managed platform also reduced the risk of private-data leakage from the real-world corpus, because participants trained and benchmarked through controlled workflows rather than downloading the full hidden dataset.

4.7 Tracks 7 and 8 Evaluation

Tracks 7 and 8 were optional out-of-domain Track 3 leaderboards. Track 7 evaluated FETV traffic-violation understanding from fisheye cameras, while Track 8 evaluated PSI-VQA pedestrian situated-intent questions. Both used VQA- and reasoning-oriented scoring related to the Track 3 protocol [9, 67] and reported final scores separately from the main TAR mean, making transfer robustness visible under fisheye geometry and socially ambiguous pedestrian-intent questions.

Award-candidate teams were required to provide reproducible code and models, and the paper-review process encouraged accepted workshop papers to report the final team names and leaderboard values exactly as shown on the evaluation system. This requirement was especially important for tracks with hidden tests or managed data access, where reproducibility depends on both the submitted model and a clearly documented training and inference pipeline.

Table 1: Track 1 leaderboard entries connected to accepted papers.

Rank	Team ID	Team	3D HOTA	Paper
1/15 (2/30)	289	EVA	56.5447	[66]; online=true
2/15 (3/30)	34	SKKU-AL-T1	52.0118	[55]; online=true
3/15 (4/30)	130	Playbox	38.0105	[53]; online=true
4/15 (5/30)	4	QDTers	34.1845	[29]; online=true
5/15 (6/30)	133	TU-YMLab	25.9712	[21]; online=true

Table 2: Track 2 leaderboard entries connected to accepted papers.

Rank	Team ID	Team	S2	Paper
1/22 (1/44)	47	Latent Painter - UTE	60.0853	[11]; VL-JEPA
2/22 (2/44)	24	UIT - Kitchen	57.3307	[38]; VQA
3/22 (4/44)	266	KZ6	56.7949	[23]; Qwen-3-VL-8B
8/22 (14/44)	127	Team KODE	55.4679	[27]; Qwen

5 Challenge Results

Tables 1–7 summarize accepted-paper entries with identifiable evaluation IDs on the 2026 leaderboards. The tables are not intended to replace the full public evaluation system; instead, they connect leaderboard outcomes to accepted workshop papers that describe the corresponding methods. Public ranks and general ranks are both reported because the public leaderboard captures award-eligible entries, while the general leaderboard captures broader participation.

5.1 Summary for the Track 1 Challenge

Track 1 attracted methods that combined object detection, camera calibration, 2D-to-3D lifting, and global multi-camera association. The leading accepted systems relied on strong RGB detectors and then used warehouse geometry to reduce cross-view ambiguity. Because the hidden test set did not expose depth, the best submissions treated depth and 3D layout as learned or calibrated priors rather than as direct test-time inputs.

The accepted Track 1 papers in Table 1 show several common design choices. EVA [66] led the public leaderboard among accepted papers with geometry-aware tracking. SKKU-AL-T1 [55] and Playbox [53] also performed strongly with online RGB-only pipelines, showing that streaming constraints remain compatible with competitive detection and association. Other papers explored collaborative 2D-3D tracking [29] and uncertainty-aware observation construction [21].

5.2 Summary for the Track 2 Challenge

Track 2 measured transfer from synthetic Digital Twin WTS data to real WTS videos. The strongest systems used VLM backbones, but the accepted papers also show that off-the-shelf prompting was rarely enough. Teams improved transfer by decomposing the task into scene parsing, phase recognition, caption generation, VQA answering, and answer calibration.

Table 3: Track 3 leaderboard entries connected to accepted papers.

Rank	Team ID	Team	Mean	Paper
1/27 (1/76)	25	Stellarview AI	0.6788	[26]; Qwen 3.5
- (2/76)	45	UOB&UW Team	0.6779	[34]; Qwen3VL-8B
2/27 (5/76)	60	FPT AI Vision	0.6703	[20]; Qwen3-VL-8B
3/27 (7/76)	12	Smart Vision	0.6669	[57]; Qwen
10/27 (26/76)	30	UWIPL ETRI	0.6185	[52]; Qwen
15/27 (36/76)	122	OptimAI	0.5880	[17]; Qwen
16/27 (41/76)	139	MR-CAS	0.5780	[31]; GPT
24/27 (55/76)	277	Korea Drive	0.4256	[25]; Qwen3

The Track 2 results in Table 2 suggest that separating scene structure from behavior helps narrow the synthetic-to-real gap. The leading accepted entries used decoupled semantics, V-JEPA style visual features, state-bridging strategies, and modular spatial grounding [11, 23, 27, 38]. Methods that represented relative position, pedestrian visibility, vehicle motion, and environmental context were better aligned with the task than generic captioning pipelines.

5.3 Summary for the Track 3 Challenge

Track 3 was the largest reasoning leaderboard in the 2026 challenge. It required systems to answer heterogeneous questions about anomalous traffic events and to produce explanations that match visual evidence. The ranking emphasized not only timestamp prediction, binary or multiple choice correctness, but also semantic and causal reasoning soundness.

As shown in Table 3, Stellarview AI [26] led both the public and general TAR ranking among accepted papers. Other strong submissions used evidence-driven chained reasoning, unified multi-leaderboard agents, metric matching, shared event memory, and structured question routing [20, 31, 34, 45, 57]. The accepted papers indicate a shift from simple VLM prompting toward agentic pipelines that first extract visual evidence, then match it to a task-specific answer format.

5.4 Summary for the Track 4 Challenge

Track 4 produced high retrieval scores, indicating rapid progress on the PAB synthetic-to-real setting. At the same time, the concentration of strong mAP values made the track sensitive to careful data-use policy and reproducibility checks. The top methods combined strong text-image embeddings with action-specific reranking and hard-negative handling.

Table 4 lists Xiilab.AIplex [47] as the top accepted public/general entry in the final decision sheet. Other accepted papers explored late consensus, heterogeneous VLM ensembling, embedding prediction, cross-encoder reranking, global assignment, and action-aligned retrieval [13, 15, 41, 49, 56, 61, 62]. The track highlights a broader lesson for text-person search: robust action semantics must be learned together with identity, clothing, and scene context.

Table 4: Track 4 leaderboard entries connected to accepted papers.

Rank	Team ID	Team	mAP	Paper
1/28 (3/47)	59	Xiilab.AI ex	99.3020	[47]
3/28 (6/47)	9	hiensumi	98.3535	[15]
7/28 (9/47)	27	VGU AI LAB	95.4078	[49]
8/28 (11/47)	97	SMART Lab	94.7815	[13]
11/28 (14/47)	76	HCMUS_4CentralVN	93.6715	[61]
13/28 (15/47)	93	SelabHuman	93.6577	[41]
15/28 (22/47)	64	GenAI4E	90.9236	[62]
18/28 (31/47)	29	EMBIA	84.2509	[56]

Table 5: Track 5 leaderboard entries connected to accepted papers.

Rank	Team ID	Team	Final	Paper
1/12 (2/18)	209	Qyn	76.4866	[14]
2/12 (3/18)	78	SSUPER	76.0385	[30]
3/12 (4/18)	47	Latent Painter - UTE	75.4302	[42]
4/12 (6/18)	83	CHTTL_A30	74.0544	[12]
5/12 (7/18)	39	VGU_ai_lab	73.3037	[28]

5.5 Summary for the Track 5 Challenge

Track 5 moved the challenge from recognizing or explaining traffic scenes to generating plausible futures. The leaderboard rewarded systems that maintained scene continuity while following textual descriptions of future behavior. This made the track a natural test of video diffusion models, world models, and text-conditioned forecasting pipelines.

The accepted systems summarized in Table 5, including Qyn [14], SSUPER [30], Latent Painter - UTE [42], CHTTL_A30 [12], and VGU_ai_lab [28], combined history-frame conditioning with language-guided future descriptions, diffusion or world-model priors, and metric-aware selection. The close scores suggest rapid progress, while the metric suite shows that safety-relevant semantic consistency remains hard to capture with one number.

5.6 Summary for the Track 6 Challenge

Track 6 emphasized a deployment problem that is often hidden by conventional object-detection benchmarks: a detector tuned for one camera network can lose accuracy when moved to another city with different viewpoints, object scales, weather, compression artifacts, and traffic composition. The Hafnia platform made this setting more realistic by using managed training and evaluation workflows, so teams had to improve domain robustness without directly extracting the full real-world corpus. This makes the track closer to practical cross-site deployment, where data governance and distribution shift are handled together.

The accepted papers in Table 6 show several complementary approaches to cross-city detection. SKKU-AL-T1 [50] led the accepted-paper entries with a strategy focused on pre-training and augmentation for zero-shot transfer. BIT-ODL [24] used evidence-conditioned multi-source pretraining, while BK2 TheFuture [51] emphasized targeted augmentation and class-aware inference.

Table 6: Track 6 leaderboard entries connected to accepted papers.

Rank	Team ID	Team	mAP	Paper
1/25 (1/35)	34	SKKU-AL-T1	0.4753	[50]
2/25 (2/35)	265	BIT-ODL	0.4281	[24]
4/25 (5/35)	261	BK2TheFuture	0.4169	[51]
- (7/35)	315	lyx	0.4060	[58]

The lyx submission [58] explored two-stage fusion and scale-aware geometric refinement. Although the final mAP values are lower than the retrieval scores in Track 4, this should be interpreted in light of the task: Track 6 evaluates object localization, class recognition, and confidence ranking under cross-city shift on real imagery. Across the accepted methods, the recurring theme is that detector architecture alone is not sufficient. Successful systems also manage resolution, category imbalance, source-domain bias, validation under distribution shift, and confidence calibration for target-city scenes.

5.7 Summary for the Track 7 and Track 8 OOD Challenges

Tracks 7 and 8 extended the Track 3 reasoning task into two out-of-domain settings. Track 7 used FETV traffic-violation understanding from fisheye cameras, where wide-angle projection changes object shape, direction cues, lane geometry, and motion interpretation. Track 8 used PSI-VQA for pedestrian situated-intent reasoning, where answers depend on subtle social signals, occlusion, ambiguity, and the viewpoint of an automated-driving agent. These tracks are therefore not merely auxiliary leaderboards; they test whether traffic reasoning systems transfer from in-domain CCTV anomaly clips to different sensing geometries and interaction questions.

Table 7 summarizes accepted-paper entries on both OOD leaderboards. UniTraffic [52] ranked first among accepted public entries on both settings, suggesting that evidence-centric agentic reasoning can transfer across anomaly, violation, and intent tasks. UniTraffic-Agent [31] and Korea Drive [25] also performed strongly with unified or task-routed video-language reasoning. TAU-Agent [34] appears in the general rankings and reflects another design pattern: retrieval-augmented reasoning can help reuse event evidence across related but shifted traffic-understanding tasks. The Track 7 scores are close among the top accepted entries, indicating that fisheye violation understanding remains sensitive to spatial evidence extraction. Track 8 has a wider score range, consistent with the added difficulty of socially ambiguous pedestrian-intent questions. Together, these OOD leaderboards complement TAR by measuring robustness, not only in-domain reasoning accuracy.

6 Discussion and Conclusion

Across a decade, the AI City Challenge has grown from vehicle detection, classification, tracking, and counting to broader urban intelligence. In 2026, 325 teams

Table 7: Track 7 and Track 8 OOD leaderboard entries connected to accepted papers.

Track	Rank	Team ID	Team	Final Paper	
Track 7: FETV traffic-violation understanding					
7	1/8 (1/15)	30	UWIPL_ETRI	0.4891	[52]
7	2/8 (3/15)	139	MR-CAS	0.4884	[31]
7	3/8 (5/15)	277	Korea Drive	0.4634	[25]
7	- (12/15)	45	UOB&UW Team	0.3998	[34]
Track 8: PSI-VQA pedestrian situated-intent reasoning					
8	1/7 (2/15)	30	UWIPL_ETRI	70.6397	[52]
8	- (5/15)	73	University of Washington	67.9275	[34]
8	4/7 (8/15)	139	MR-CAS	64.4161	[31]
8	5/7 (9/15)	277	Korea Drive	57.0400	[25]

from 26 countries and regions evaluated systems across six primary tracks and two OOD leaderboards: multi-camera 3D perception, synthetic-to-real safety understanding, anomaly reasoning, text-based search, generative forecasting, and cross-city detection. The track portfolio reflects how the field has moved from isolated CV predictions toward systems that combine geometry, language, temporal evidence, generation, and domain adaptation.

The 2026 results also show that progress is uneven in productive ways. Some teams achieved strong retrieval and reasoning scores with foundation-model pipelines, while the cross-city detection, RGB-only 3D perception, video forecasting, and OOD reasoning settings exposed remaining gaps in transfer, calibration, and reproducibility. This is a useful outcome for a benchmark: the best submissions identify practical solution patterns, while the remaining failures define research directions for future editions. We expect subsequent challenges to emphasize reproducible end-to-end systems, more explicit OOD testing, privacy-preserving evaluation platforms, and interaction with related programs such as IARPA Video LINC S [22].

Across the accepted papers, three patterns are especially visible: many high-ranking entries used modular pipelines, the strongest Sim2Real and cross-city methods treated domain shift as a first-class constraint, and the multimodal tracks showed that language is useful only when grounded in visual evidence. Together, these patterns connect the 2026 tracks into a shared research agenda for deployable urban AI.

Acknowledgments

Rama Chellappa was supported by the VideoLINC S program. This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 56000026C0026. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

1. Abduljawad, A., S, M.S., Shaik, N.S., Chang, M.C., Hsieh, J.W., Gochoo, M.: FETV: Traffic violation report generation from fisheye camera systems. In: ECCV Workshops. Malmö, Sweden (2026)
2. AI City Challenge: 2026 evaluation system. <https://www.aicitychallenge.org/2026-evaluation-system/> (2026), accessed 2026-08-05
3. AI City Challenge: 2026 track 1: Multi-camera 3d perception (sim2real). <https://www.aicitychallenge.org/2026-track1/> (2026), accessed 2026-08-05
4. AI City Challenge: 2026 track 2: Transportation safety understanding and captioning (sim2real). <https://www.aicitychallenge.org/2026-track2/> (2026), accessed 2026-08-05
5. AI City Challenge: 2026 track 3: Anomalous events in transportation. <https://www.aicitychallenge.org/2026-track3/> (2026), accessed 2026-08-05
6. AI City Challenge: 2026 track 4: Text-based person anomaly search (sim2real). <https://www.aicitychallenge.org/2026-track4/> (2026), accessed 2026-08-05
7. AI City Challenge: 2026 track 5: Generative traffic video forecasting. <https://www.aicitychallenge.org/2026-track5/> (2026), accessed 2026-08-05
8. AI City Challenge: 2026 track 6: Cross-city object detection (milestone project haffnia). <https://www.aicitychallenge.org/2026-track6/> (2026), accessed 2026-08-05
9. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: Visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2425–2433 (2015)
10. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. pp. 65–72 (2005)
11. Bui, N.H.T., Vo, T.N., Nguyen, T.T.G., Bui, H.D.: Sim-to-real traffic scene understanding by decoupling semantics from caption generation with V-JEPA. In: ECCV Workshops. Malmö, Sweden (2026)
12. Chu, Y.H., Lu, B.A., Du, S.W., Chen, W.C.: LLM-guided kinematic caption fusion for traffic video forecasting with frozen world foundation models. In: ECCV Workshops. Malmö, Sweden (2026)
13. Denteh, E.K.O., Kyem, B.A., Danyo, A., Asamoah, J.K., Dzinyela, R., Owusu-Ansah, S., GYIMAH, D.A., Aboah, A.: Synthetic-to-real text-based person anomaly search via multi-model retrieval fusion and cross-encoder reranking. In: ECCV Workshops. Malmö, Sweden (2026)
14. Dinh, Q.M., Doan, T.K.: CosmosAlign: Adapting a world foundation model for generative traffic video forecasting. In: ECCV Workshops. Malmö, Sweden (2026)
15. Duy, H.P., Gia, H.N.P., Tran, B., Nguyen, T., Do, T., Ngo, T.D., Duy-Dinh Le, V.H., Satoh, S.: Protocol-aware global assignment for text-based person anomaly search. In: ECCV Workshops. Malmö, Sweden (2026)
16. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision* **88**, 303–338 (2010)
17. Gankhuyag, G., Yoo, J., Park, J.M., Lee, S., Son, H., Min, K.: Motion trails as visual prompts: A vision-language pipeline for traffic anomaly reasoning. In: ECCV Workshops. Malmö, Sweden (2026)

18. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: CLIPScore: A reference-free evaluation metric for image captioning. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. pp. 7514–7528 (2021)
19. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: *Advances in Neural Information Processing Systems* (2017)
20. Hoang, V.A.: Metric-matched test-time selection for traffic anomaly video question answering. In: *ECCV Workshops*. Malmö, Sweden (2026)
21. Hsu, Y.S., Zielewski, M., Huang, C.Y.: Uncertainty-guided multi-camera observation construction for online tracking. In: *ECCV Workshops*. Malmö, Sweden (2026)
22. IARPA: Video lincs: Video linking and intelligence from non-collaborative sensors. <https://www.iarpa.gov/research-programs/video-lincs> (2024), accessed 2026-08-07
23. Jeong, J., Kim, S., Gankhuyag, G.: Modular spatial grounding for sim2real traffic-safety VQA. In: *ECCV Workshops*. Malmö, Sweden (2026)
24. Jia, Z., Wang, J., Zou, Y., lili: EC-DEIM: Evidence-conditioned multi-source pre-training for cross-city fine-grained object detection. In: *ECCV Workshops*. Malmö, Sweden (2026)
25. Kim, H.: Task-routed video-language reasoning across traffic domains: Korea drive at ai city challenge 2026. In: *ECCV Workshops*. Malmö, Sweden (2026)
26. Kulkarni, M.: Reason or recite: Complementary specialists with an LLM judge for traffic anomaly reasoning. In: *ECCV Workshops*. Malmö, Sweden (2026)
27. Kyem, B.A., Dzinyela, R., Asamoah, J.K., Denteh, E., Danyo, A., Owusu-Ansah, S., Asiedu, D.G., Aboah, A.: StateBridge: Phase- and track-grounded reasoning for synthetic-to-real traffic video captioning and VQA. In: *ECCV Workshops*. Malmö, Sweden (2026)
28. Le, K.M., Pham, H.D.T., Danh, L.T., Le, N.T., Ngo, H.A., Tran, P.H.V., Le, S.N.M., Nghia, N.T., Cam, T.T.T., Nguyen, H.M.N., Nguyen, C.T.: GeoRoute: Geometry-aware hybrid inference for traffic future-frame prediction. In: *ECCV Workshops*. Malmö, Sweden (2026)
29. Le, V.M., Group, D.L.H.V., Hanoi, V.N.U., Institute, H.N.V.A., Hanoi, V.N.U., Tran, H., Tran, L.Q.: CoTrack3D: A collaborative 2d-3d framework for indoor multi-target multi-camera tracking. In: *ECCV Workshops*. Malmö, Sweden (2026)
30. Lee, J., Kim, D., heum Kim, S.: Generative traffic video forecasting with paired future descriptions via parameter-efficient world model adaptation. In: *ECCV Workshops*. Malmö, Sweden (2026)
31. Li, P., Xu, Q., Bao, S., Jiang, Y., Huang, Q.: UniTraffic-Agent: Unified traffic video reasoning for ai city challenge 2026 track 3 with two out-of-domain evaluations. In: *ECCV Workshops*. Malmö, Sweden (2026)
32. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: *Text Summarization Branches Out*. pp. 74–81 (2004)
33. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: *European Conference on Computer Vision*. pp. 740–755 (2014)
34. Lin, Y., Shi, Y., Lockyer, S., Madabushi, H.T., Evans, A.N., Li, W., Wang, Y., Zhang, N.: TAU-Agent: An agentic retrieval-augmented framework for traffic anomaly understanding. In: *ECCV Workshops*. Malmö, Sweden (2026)
35. Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixe, L., Leibe, B.: Hota: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision* (2021)

36. Milestone Systems: Hafnia dataset: Eeccv cross city object detection dataset. <https://hafnia.milestonesys.com/datasets/ad88fd6e-d608-46e9-8d23-fbe94734cf19> (2024), part of the Hafnia project, version 0.0.1
37. Milestone Systems: Hafnia Training-as-a-Service. <https://hafnia.milestonesys.com/> (2026), accessed: 2026-08-13
38. Minh, N.D., Tuan, M.N.H., Kim, L.H.L., Minh, L.V.N., Minh, H.P.D., Tran, B., Do, T., Thanh, Ngo, D., Le, D.D., Satoh, S.: A semantic state bridge for Sim2Real traffic-safety VQA and captioning. In: ECCV Workshops. Malmö, Sweden (2026)
39. MLCG Lab: SynWTS dataset. <https://huggingface.co/datasets/mlcglab/synwts> (2026), AI City Challenge Track 2 synthetic traffic-safety dataset
40. Naphade, M., Wang, S., Anastasiu, D.C., Tang, Z., Chang, M.C., Yao, Y., Zheng, L., Rahman, M.S., Arya, M.S., Sharma, A., Feng, Q., Ablavsky, V., Sclaroff, S., Chakraborty, P., Prajapati, S., Li, A., Li, S., Kunadharaju, K., Jiang, S., Chellappa, R.: The 7th AI city challenge. arXiv preprint arXiv:2304.07500 (2023)
41. Nguyen, T.K., Vo, T.N., Tran, M.T.: Action-aligned retrieval with pairwise multimodal reranking for text-based person anomaly search. In: ECCV Workshops. Malmö, Sweden (2026)
42. Nguyen, T.T.G., Vo, T.N., Bui, N.H.T., Bui, H.D.: JEPA guided diffusion: Predictive vision-language conditioning for generative traffic forecasting. In: ECCV Workshops. Malmö, Sweden (2026)
43. NVIDIA: PhysicalAI-smartspaces dataset. <https://huggingface.co/datasets/nvidia/PhysicalAI-SmartSpaces> (2026), AI City Challenge Track 1 dataset
44. NVIDIA: PhysicalAI-traffic-anomaly-reasoning dataset. <https://huggingface.co/datasets/nvidia/PhysicalAI-Traffic-Anomaly-Reasoning> (2026)
45. Pandey, A.K., Chang, M.C.: TRACE-TAR: Reusable video evidence and shared event memory for transportation anomaly reasoning. In: ECCV Workshops. Malmö, Sweden (2026)
46. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
47. Park, J., Lee, J.: Preserving priors, resolving collisions: Complementary reranking and injective assignment for text-based person re-identification (sim2real). In: ECCV Workshops. Malmö, Sweden (2026)
48. Patil, D., Kachhadiya, R., Vattuone, A., Liang, H., Li, J., Wang, Y., Tang, Z., Kumar, A., Kong, Q., Anastasiu, D.C.: SynWTS: Multi-view digital-twin dataset for synthetic-to-real traffic safety understanding. In: ECCV Workshops. Malmö, Sweden (2026)
49. Pham, H.D.T., Tran, P.H.V., Duc, T.M., Le, S.N.M., Khang, L.M., Vo, H., Phung, M.C., Nguyen, H.M.N., Nguyen, C.T.: FaLCon: Facet-anchored retrieval with late consensus for sim2real text-based person anomaly search. In: ECCV Workshops. Malmö, Sweden (2026)
50. Pham, L.H., Ho, Q.P.N., Nguyen, H.H., Tran, D.N.N., Huynh, N.D.M., Le, C.Q., Nguyen, H.K., Jeon, H.M., Tran, C.D., Phan, S.H., Vu, D.K., Khanh, T.L.B., Jeon, J.W.: Rethinking pre-training and augmentation for zero-shot cross-city object detection. In: ECCV Workshops. Malmö, Sweden (2026)
51. Phat, P.H., Bang, C.T., Vu, K.D., Nguyen, D.D.: Cross-city traffic object detection: A targeted augmentation and class-aware inference approach. In: ECCV Workshops. Malmö, Sweden (2026)
52. Shangguan, J., Kim, S., Huang, H.W., Sun, W.C., Lu, J., Kim, P., Kim, K.J., Hwang, J.N.: UniTraffic: Evidence-centric agentic video reasoning across traffic anomalies, violations, and intent. In: ECCV Workshops. Malmö, Sweden (2026)

53. Shrestha, P., Nakayama, H., Scott, A.: Online multi-camera 3d tracking via id prediction over recurrent sparse queries. In: ECCV Workshops. Malmö, Sweden (2026)
54. Tang, Z., Wang, S., Anastasiu, D.C., Chang, M.C., Sharma, A., Kong, Q., Kobori, N., Gochoo, M., Batnasan, G., Otgonbold, M.E., Alnajjar, F., Hsieh, J.W., Kornuta, T., Li, X., Zhao, Y., Zhang, H., Radhakrishnan, S., Jain, A., Kumar, R., Murali, V.N., Wang, Y., Pusegaonkar, S.S., Wang, Y., Biswas, S., Wu, X., Zheng, Z., Chakraborty, P., Chellappa, R.: The 9th AI city challenge. arXiv preprint arXiv:2508.13564 (2025)
55. Tran, D.N.N., Huynh, N.D.M., Le, C.Q., Nguyen, H.K., Pham, L.H., Nguyen, H.H., Ho, Q.P.N., Khanh, T.L.B., Tran, C.D., Vu, D.K., Phan, S.H., Jeon, H.M., Jeon, J.W.: Syn2RealTrack: Bridging the gap between synthetic and real world dataset for online multi-view multi-target tracking. In: ECCV Workshops. Malmö, Sweden (2026)
56. Traore, A., Couturier, A., Hervet, E.: SCOUT: Sim-to-real text-based person retrieval by embedding-space prediction over frozen video features. In: ECCV Workshops. Malmö, Sweden (2026)
57. Truong, D., Dinh, K.D., Hoang, M.H., cuong tien nguyen, Cao, H.T., Luong, T.V.: EDCR: Evidence-driven chained reasoning for unified traffic anomaly understanding. In: ECCV Workshops. Malmö, Sweden (2026)
58. Tsai, C.M., Li, Y.X., Hsieh, J.W., Chang, M.C.: Decoupled two-stage fusion with scale-aware geometric refinement for cross-city object detection. In: ECCV Workshops. Malmö, Sweden (2026)
59. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric and challenges. arXiv preprint arXiv:1812.01717 (2018)
60. Vedantam, R., Zitnick, C.L., Parikh, D.: CIDEr: Consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4566–4575 (2015)
61. Vo-Lan, T., Nguyen-Ngoc, P., Khanh, T.Q., Thanh, V.T.: STAR: Action-aware two-stage retrieval for text-based person anomaly search. In: ECCV Workshops. Malmö, Sweden (2026)
62. Vu, H.A., Cam, T.T.T., Toan, L.N.T., Vo, H., Hieu, D.T., Pham, H.D.T., Le, K.M., Nguyen, H.M.N.: Heterogeneous vision-language ensemble with disagreement-aware reranking for text-based person anomaly retrieval. In: ECCV Workshops. Malmö, Sweden (2026)
63. Wang, S., Lu, Y.C., Tang, Z., Zhang, H., Anastasiu, D.C., Tian, R.: PSI-VQA: Evaluating multimodal LLM social reasoning under ambiguous pedestrian intent in automated driving. In: ECCV Workshops. Malmö, Sweden (2026)
64. Wang, S., Anastasiu, D.C., Tang, Z., Chang, M.C., Yao, Y., Zheng, L., Rahman, M.S., Arya, M.S., Sharma, A., Chakraborty, P., Prajapati, S., Kong, Q., Kobori, N., Gochoo, M., Otgonbold, M.E., Alnajjar, F., Batnasan, G., Chen, P.Y., Hsieh, J.W., Wu, X., Pusegaonkar, S.S., Wang, Y., Biswas, S., Chellappa, R.: The 8th AI city challenge. arXiv preprint arXiv:2404.09432 (2024)
65. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing* **13**(4), 600–612 (2004)
66. Wu, X., Fang, H., Xi, X., Tang, Y., Xu, D., Bai, X., Liang, D.: Aligning learned spatial priors: Gauge canonicalization for multi-camera 3d perception under sim-to-real transfer. In: ECCV Workshops. Malmö, Sweden (2026)

- 67. Zhang, H., Zhao, Y., Bhat, Z.P., Tang, Z., Praveen, V., Murali, V.N., Anastasiu, D.C., Kornuta, T.: From detection to understanding: Tar and tar-bench for multi-task traffic anomaly reasoning (2026), <https://arxiv.org/abs/2608.10317>
- 68. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018)