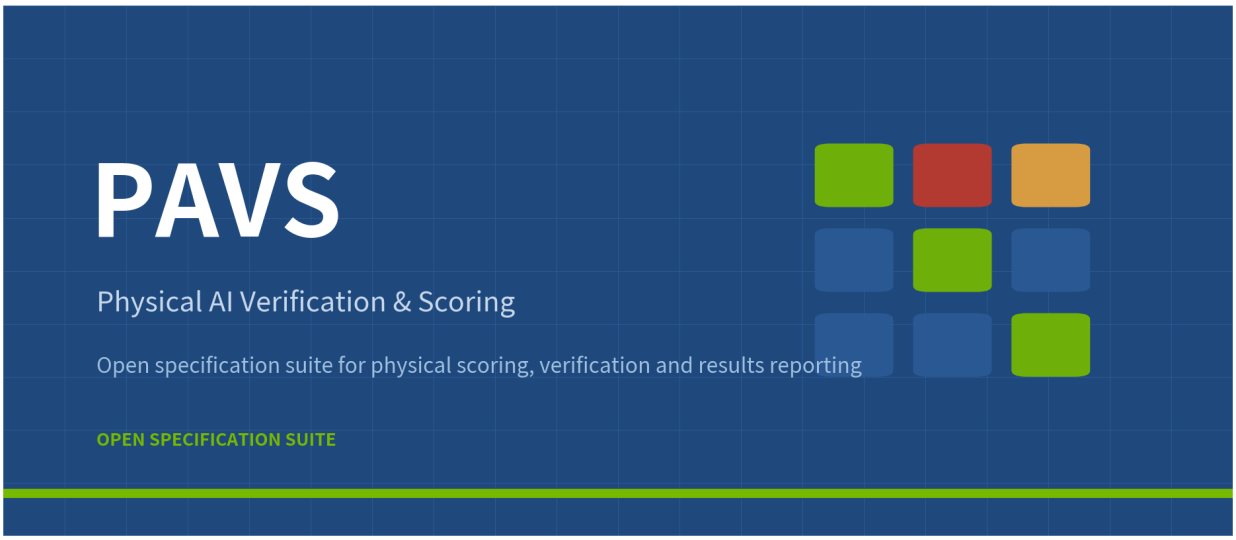


Part 0 — White Paper (informative) · English translation

Issued edition v1.5 · 29 October 2026

Published by ImmersiveCast Co., Ltd. · open specification at the proposal stage (not a standard of a formal standards body)

English translation — the Korean text is authoritative · immersivecast.com/pavs

Contents

Summary

0. About this document

1. Background — the precedent of automated driving and the standards gap in Physical AI

1.1 The most mature precedent of simulation scoring becoming institutionalized
1.2 Robots have no equivalent
1.3 The obligation to evaluate is being imposed now

2. The problem — a single success-rate figure is not proof of capability

3. The answer — the five links of the chain of trust in scoring

4. Links 1 and 2 — items and measurement

4.1 Items: scenarios in a self-contained domain language
4.2 Measurement — distinguishing the existence of a value from a measurement

5. Link 3 — verdict: not-measured is a third state

6. Link 4 — verification: scoring the scorer

7. Link 5 — reporting: performance boundaries, not scores

8. Origin of the clauses — the place of the Casebook (Part 2)

9. Conformance — requirements that come with test methods

9.1 A test method for every requirement
9.2 Conformance classes
9.3 Reference implementation
9.4 The party that wrote the specification is tested first

10. Position in the ecosystem

11. Open specification roadmap

References

Appendix A (informative) — Correspondence between automated-driving precedents and this specification

Appendix B (informative) — Correspondence table of reference standards and KS

B.1 Position on interfaces with Korean standards

Appendix C (informative) — Wording boundaries for citation

Appendix D (informative) — Glossary (plain-language explanations)

Appendix E (informative) — Self-test checklist (Annex A, 72 assertions)

Revision notice — v1.4 → v1.5 (2026-10-20 · errata correction)

Revision notice — v1.3 → v1.4 (2026-10-20 · same-day textual correction)

Revision notice — v1.2 → v1.3 (2026-10-20)

Revision notice — v1.1 → v1.2 (2026-10-19)

Revision notice — v1.0 → v1.1 (2026-10-02)

Inquiries

English edition — translation. This is the English translation of the Korean edition v1.5 (2026-10-20). The Korean text is authoritative; where the two differ, the Korean text prevails. Assertion identifiers, clause numbers, case identifiers, versions, dates and figures are identical to the Korean edition.

PAVS (Physical AI Verification & Scoring) — an open specification for physical scoring, verification and results reporting. **Document status**: Issued edition v1.5 · 2026-10-29 (errata correction — modal forms of requirement quotations · terms · Appendix D) · previous issue v1.4 2026-10-20 · first public release v1.0 2026-09-09 · corresponding specification Part 1 v1.5 (72 test assertions · 8 ★core · requirements identical to v1.4) · published at immersivecast.com/pavs. This document (Part 0) is **informative** (explanatory) in its entirety. The normative clauses are in Part 1. The changes in this edition are given in the revision notice at the end of the document. **Conflict-of-interest notice**: The publisher, ImmersiveCast Co., Ltd., is not a certification body or an accreditation body, and is the same as the author of the implementation (Part 4) that is evaluated under this specification. Neutrality is assured only by setting no requirement favorable to any particular product in the body of the specification. **Drafting principles**: This White Paper describes the target state toward which the specification is directed. Figures are included only where the argument requires them, and cases are cited only as the grounds from which clauses arose. **Citation convention**: External facts are given together with their source and date. The list of sources is given in the Bibliography.

Summary

■ Background

- Automated driving has established scenario domain languages and simulation verification as the formal methodology of international regulations and assessment protocols (UNECE R157, NATM and Euro NCAP)
- Robots and Physical AI have no equivalent — existing robot standards presuppose physical testing, and for humanoids the development of the relevant safety standards is at the initiation stage (publication expected to take several years — industry estimate)
- The ‘obligation’ to evaluate is already being imposed by regulatory regimes ahead of the methodology (EU Machinery Regulation in full application from 2027-01, Korea's AI Framework Act in force from 2026-01-22, Outdoor Mobile Robot Operational Safety Certification in force since 2023-11, national Physical AI strategy published 2026-07)

■ Problem statement

- Results reporting centered on a single success-rate figure and demonstration videos is insufficient as proof of capability — vulnerability to changes in conditions (collapses from 90%+ to 0% reported), the statistical powerlessness of customary trial counts, and the inability to determine autonomy
- If the reliability of the scoring tools and models themselves is unverified, even high scores are meaningless — most records of incidents actually observed are of the type that is ‘wrong silently, in a plausible-looking form, with no error indication’

■ Proposal

- Proposes PAVS, an open specification that specifies as requirements the 5-link chain of trust in scoring (item–measurement–verdict–evidence–reporting; an item is an individual element of scoring)
- Defines all requirements as 72 test assertions (a single-sentence condition whose truth or falsity can be determined; see the glossary in Appendix D) and specifies test methods for all 72 (36 by injection, 33 by audit and 3 by demonstration) — conformance is

determined from the records of tests performed by these procedures, not by document review alone
- Every clause is mapped to its grounds of origin — an incident from in-house operations, a failure record in the published literature, or the acceptance conditions of a precedent (correspondence table: Part 1 Annex B; incident narratives: Part 2)
- The publisher applies it first to its own implementation — the self-test of 2026-09-04 (against Part 1 v1.2, Class C1) found 7 of 72 met and 3 of the 8 ★core assertions met; the result is indicated as “PAVS-aligned”, and every assertion met and not met is disclosed (9.4, Appendix E)

■ Structure and use

- Part 0 White Paper (design philosophy), Part 1 Specification (requirements and conformance), Part 2 Casebook (commentary), Part 3 Reference Implementation & Reproduction, and Part 4 Solution Profile (an example implementation — unpublished, to be published next)
- Uses: the methodology layer for building verification systems, for reference in procurement requirements and for the design of evaluation schemes — a staged path of open specification + reference implementation → consortium → formal standardization is envisaged

0. About this document

PAVS is an open specification suite that defines the **methods of scoring in simulation, verifying that scoring and reporting the results** for the capability and safety of Physical AI — artificial intelligence systems that perceive, move and manipulate in the physical world. This White Paper examines the background that makes this specification necessary and the global landscape, and presents the design philosophy and structure of the specification. The suite consists of five parts — the Specification is neutral with respect to products and organizations, and the name of an implementation is given only in the profile (Part 4). Part 4 has not yet been published and will be published next.

Part 0 White Paper	Part 1 Specification	Part 2 Casebook	Part 3 Reference Implementation	Part 4 Solution Profile
Need · design philosophy (informative)	Terms · 72 requirements · conformance (normative)	39 cases from which clauses arose (commentary)	Mechanisms · 3 demonstrations · reproduction package (informative · published)	Configuration · boundaries · delivery form of one implementation (informative · unpublished, next)

The current edition of each part is determined by the document list at the publication location (immersivecast.com/pavs).

Normative vocabulary: Part 1 uses a 4-level vocabulary of shall (requirement)/should (recommendation)/may (permission)/can (possibility), and uses must only when referring to legal constraints external to the document (following the convention of ISO/IEC Directives Part 2). The text of Part 1 is entirely normative except for NOTEs, EXAMPLEs and clauses marked informative. The remaining parts, including this White Paper, are all explanatory documents.

Honest indication of status: PAVS is an **open specification** at the proposal stage. It is not a standard of a formal standards body (ISO, IEC or KS), and in this document expressions such as "standard scenario" and "standard item" refer to **objects defined by this**

specification. Existing formal standards (ISO 9283, ISO 10218, ISO 18646, ISO 3691-4, etc.) are cited as **reference standards**, and this does not imply conformance certification to them.

Publisher and conflict of interest: The publisher, ImmersiveCast Co., Ltd., is not a certification body or an accreditation body, and is the same as the author of the implementation (Part 4) that is evaluated under this specification. Because the party that wrote the specification and the party evaluated under it are the same, neutrality is assured only by setting no requirement favorable to any particular product in the body of the specification, and is not assured by the independence of the publisher (Part 1 12.1). Transfer to a third-party publisher is a goal of this specification and has not taken place in the current edition.

1. Background — the precedent of automated driving and the standards gap in Physical AI

> **Automated driving made simulation verification a formal methodology of the regulatory regime. For robots and Physical AI that layer is empty, and the obligation to evaluate is already being imposed.**

1.1 The most mature precedent of simulation scoring becoming institutionalized

Automated driving is the most mature precedent in which a regulatory regime has recognized the results of simulation verification. Whereas the acceptance of simulation in other industries — flight simulator qualification, computational modeling evidence for medical devices — remains at the layer of tool qualification and individual evidence, automated driving has codified scenario-based verification as the methodology for the type approval of mass-produced products.

- **UNECE R157** (2021) introduced the staged structure of "virtual simulation scenario verification → closed track → real road" into the international type approval of Level 3 Automated Lane Keeping Systems, and requires proof of the scope and validity of the simulation tool and toolchain verification (correlation with physical testing).
- **UNECE NATM** (adopted 2021-06) formalized a multi-pillar methodology consisting of a scenario catalog and 5 verification pillars (including virtual testing).
- The **Euro NCAP 2026 protocols** (press release 2025-11-26; the virtual testing protocols were released in advance in 2025-03 and took effect in 2026-01) are the first cycle to formally incorporate virtual test data into active safety (ADAS) assessment (submission of virtual test data for part of passive safety has been mandatory since 2024). The driver of adoption was that the explosion in scenario combinations (an increase of about 186% — according to interviews with Euro NCAP technical staff) made exhaustive physical testing impractical, and the condition of acceptance is the proof of simulation-to-vehicle correlation — quantitative criteria for agreement scores, metric tolerances and per-cluster satisfaction rates (some of them provisional).

Underlying this path is **a lineage of writing scenarios in a domain language**. ASAM OpenSCENARIO/OpenDRIVE made test scenarios into a language that people read and machines execute, and established a lineage in which regulatory text (the test scenarios of R157) becomes executable code.

1.2 Robots have no equivalent

The standards landscape for robots and Physical AI is different.

- **The vocabulary of measurement items already exists.** ISO 9283 precisely specifies the pose accuracy and repeatability of manipulators; the ISO 18646 series, the performance of service robots in locomotion, navigation, manipulation and the like (with new parts, such as legged locomotion and lower-limb wearable robots, under development); ISO 3691-4, the safety requirements for driverless industrial trucks; and ISO 10218:2025, industrial robot safety (with collaborative operation integrated).
- **However, these standards presuppose physical testing.** "Measure this" exists, but no standard specifying "how to reproduce this in simulation, how to trust that measurement and how to report the results" was found within the scope of the survey (2026-08, based on publicly available material) — there are conventions for confidence in software tools in the field of functional safety, but they do not address the layer of simulation scoring and reporting of robot performance and safety items.
- **Adjacent normative documents exist, and this specification inherits from them.** The ISO/IEC 9646 family established the 3-valued verdict of conformance testing (pass/fail/inconclusive) and the Implementation Conformance Statement (ICS). ISO/IEC TS 42119-2:2025 specifies the test process for AI systems; ISO/IEC 17025, the competence, uncertainty and decision rules of testing laboratories; and ASME V&V 40 and NASA-STD-7009, the credibility assessment of computational models. This specification inherits from them and adds the layer beneath them (the provenance of physical-quantity measurement in simulation and the establishment of verdicts) (Part 1 Annex D). Matters specified by predecessor standards but not yet provided for by this specification (decision rules for uncertainty intervals, risk-proportionate credibility) are also stated in the same Annex.
- **For humanoids, the standards themselves are still being made.** The IEEE humanoid standards study group (report of 2025-09) stated explicitly that existing robot safety standards cannot handle dynamically unstable walking bodies, and the development of the related ISO safety standard (dynamically stable mobile robots — including humanoids and multi-legged robots, ISO 25785-1) is at the initiation stage (initiation announced 2025-05), and the industry expects publication to take several years.

1.3 The obligation to evaluate is being imposed now

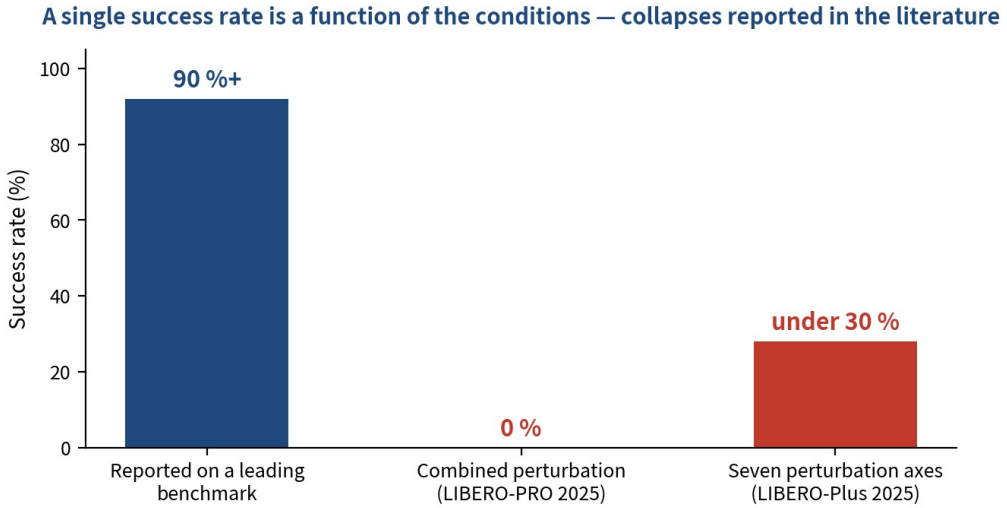
Regardless of the gap in methodology, regulatory regimes are already imposing the **obligation** to evaluate.

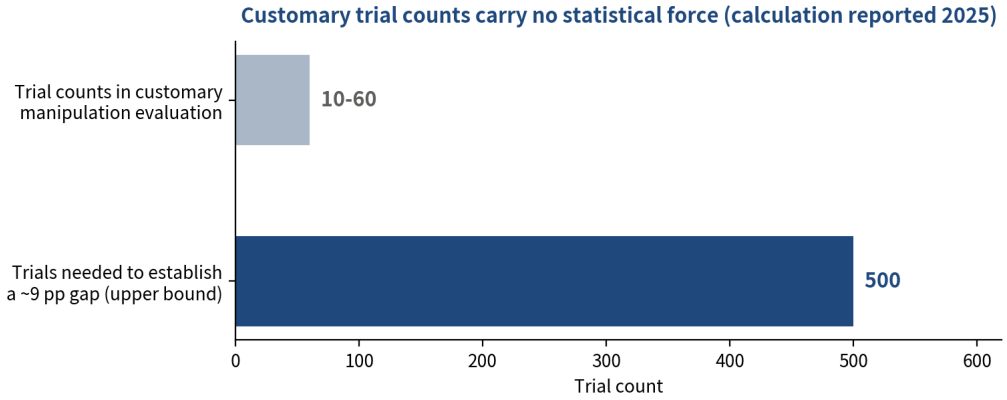
- The EU Machinery Regulation (2023/1230) applies in full from 2027-01-20 and requires third-party conformity assessment for machine-learning safety-function components whose behavior changes through learning (fully or partially self-evolving) — harmonized standards for the test methods are still being prepared.
- Korea is moving in the same direction with the AI Framework Act (“Framework Act on the Development of Artificial Intelligence and the Establishment of a Foundation for Trust”, in force from 2026-01-22), and the government has formalized that direction with the “Strategy for Securing Core Competitiveness in Physical AI” (Ministry of Science and ICT, published 2026-07-01 — 4 main pillars: data acquisition, core technologies, service diffusion and ecosystem building). The Outdoor Mobile Robot Operational Safety Certification (Intelligent Robots Act, in force since 2023-11) already operates a structure in which "the right to enter public spaces is obtained only by passing the scoring" — however, its tests are based on physical hardware (real robots), and no provision recognizing simulation results in the certification criteria was found (2026-09 survey).

- National virtual field-testing **infrastructure** is being prepared — the National Robot Test Field has specified virtual-environment (simulation) field-testing evaluation in its field-testing support platform, and has launched, as its first project, technology development of a digital-twinning framework (2024–2028). Infrastructure (sites and equipment) and **scoring methodology** (specification of items, thresholds and evidence) are different layers, and what this specification addresses is the latter — it does not compete with infrastructure investment but aligns with it.

> **NOTE** — The structure of the argument in this clause: "what to measure" (existing standards) and "the obligation to measure" (new regimes) are being put in place, but "the method of measuring, trusting and reporting through simulation" is empty. PAVS is the specification that fills that layer.

2. The problem — a single success-rate figure is not proof of capability





> **A single success-rate figure is a function of conditions, and a demonstration video does not even prove autonomy — scoring is a chain of item, measurement, verdict, evidence and reporting.**

Today the results of Physical AI are mostly reported in two forms: **demonstration videos** and **benchmark success rates**. Neither is scoring. Recent research has shown this quantitatively.

1. **A success rate is a function of conditions.** Models that scored 90% or more on a representative benchmark collapsed to a 0% success rate under combined-perturbation evaluation, and a single axis of object position displacement alone made them plunge to 0% beyond a certain displacement (LIBERO-PRO, 2025). They also "succeeded" on meaningless string instructions with the same trajectories as for normal instructions — the authors' conclusion was "memorization of action sequences", not task understanding. Results in which the success rate plunged from 95% to below 30% under only 7 axes of moderate perturbation (placement, viewpoint, initial state, lighting, etc.) have also been reported (LIBERO-Plus, 2025).

2. **Customary trial counts are statistically powerless.** The customary practice in comparing robot manipulation policies is 10–60 trials. A calculation has been presented showing that up to 500 trials are needed to establish a success-rate gap of about 9 percentage points (59% versus 68%) (2025). The practice of adding trials when results are ambiguous invalidates the statistical guarantee.

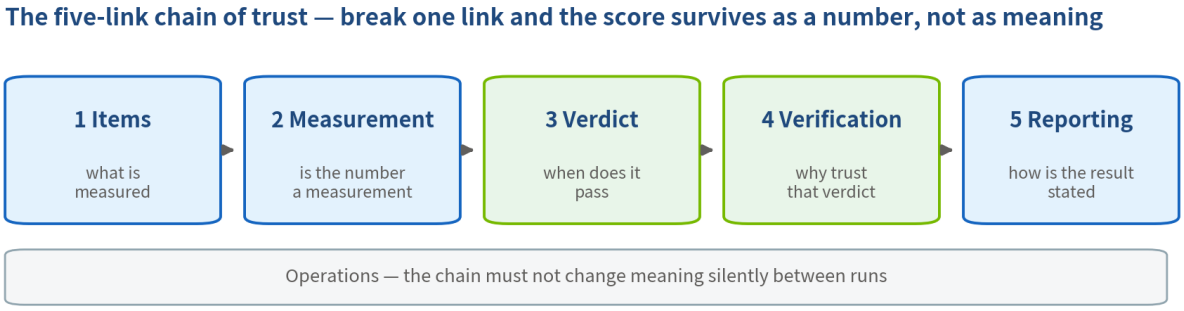
3. **Video does not even prove autonomy.** Public demonstrations of commercial humanoids have repeatedly become embroiled in teleoperation controversies (2024–2025 — some are interpretations based on circumstantial evidence without official acknowledgment by the manufacturer concerned, while in others consent to remote assistance is stated explicitly in the manufacturer's official guidance and use-approval procedure). Without logs of the operating agent, video alone cannot determine even "who did it".

4. **The scoring tool itself changes the verdict.** A cross-simulator benchmark that found by measurement that 4 physics engines differ from one another in error relative to real-world measurement and in stability on the same cloth (deformable object) manipulation task was published after peer review (RA-L 2024). A public locomotion learning framework (Humanoid-Gym) makes it the default procedure to re-verify, in another engine, a policy trained in one physics engine.

Large-scale evaluation practice in industry points in the same direction. An evaluation system equipped with large-scale blind A/B testing, control of initial conditions, rubric-based partial credit and **quality control of the evaluation itself** (measurement of re-review rates and rubric disagreement rates) was published with the explicit statement that a success rate alone cannot tell the whole story of policy performance (2025).

> **NOTE** — The direction in which the research community is converging is clear: ① moving beyond a single success rate (partial credit, failure modes) ② measurement of the reliability of scoring tools ③ statistical guarantees ④ turning perturbation and robustness into items ⑤ third-party evaluation. PAVS is an attempt to integrate this direction into **a single operable specification**; it is not an original discovery but a translation of what international research demands into an executable form.

3. The answer — the five links of the chain of trust in scoring



Trust in test results is established when five requirements are met in sequence. The structure of PAVS is those five requirements.

Trust in test results is established when five requirements are met in sequence. The structure of PAVS reflects these five requirements directly.

#	Link	Requirement	Corresponding clauses of Part 1
1	**Item**	Definition of what is measured — the clarity, executability and discriminating power of items are verified	Scenario and item requirements
2	**Measurement**	Source of values — values used in scoring are proven to be measured	Measurement provenance requirements
3	**Verdict**	Pass criteria — not-measured is treated as a third state that is neither pass nor fail. Where a mandatory item lacks measurement, it may be treated as a fail for the reason ‘absence of measurement’, and the count of not-measured results is retained and indicated separately	Verdict rule requirements
4	**Verification and evidence** (Evidence)	Grounds of the verdict — the scorer itself is verified and results are reproduced	Verification and reproducibility requirements
5	**Reporting**	Manner of describing results — the score report states its own limitations and	Results reporting requirements

		attribution	

Each link is described as a target state in the following clauses. The operational requirements (the register of failed methods and record-keeping discipline) are a cross-cutting support layer that sustains the five links, and form a separate group in Part 1 (Clause 10). Clauses 4 to 7 restate the gist of the normative clauses of Part 1, and this White Paper creates no requirements. The authoritative wording of the requirements is that of Part 1.

4. Links 1 and 2 — items and measurement

> **The substance of the specification is not a document but a set of executable scenarios (a corpus), and scoring that does not ask for the source of values goes wrong in both directions, as false passes and as false fails.**

4.1 Items: scenarios in a self-contained domain language

Test scenarios shall be described in a self-contained manner in a **domain language** that people read and machines execute. A third party shall be able to reproduce the same test from the same description alone, and scenario logic shall not be dispersed into external auxiliary code or hardcoded. This is the robot version of the practice that automated driving established with OpenSCENARIO.

The substance of the specification is not a document but a corpus. A collection that provides executable scenarios, controls and expected verdicts as one set (a corpus) is the tangible form of the specification, and the deliverables of the specification shall include a machine-readable scenario library (the convention by which ASAM includes domain model libraries as official deliverables).

Items themselves are also subject to verification. The value of an item lies not in the fact that scoring was performed but in its ability actually to separate pass from fail (discriminating power). An item that yields a pass for any input (an item like a proposition that is always true, that is, a tautology) catches no defect, and for an item that has never once yielded a fail, whether the boundary between pass and fail is correct remains unconfirmed. The specification classifies items into four classes by their test history (discriminating, constant-pass (always passes), constant-fail (always fails) and ineligible), and requires that the boundary of an item that has only ever yielded one verdict be confirmed only by a **reverse-direction control** (a scenario in which only the stimulus of the original scenario is changed to the opposite direction). The approach of changing the threshold to produce the desired verdict is not permitted.

Rated capability is confirmed by sustainability. For each rated capability published by the manufacturer (payload, speed, continuous operation, etc.), there shall be an item that confirms whether that capability is sustained under the declared conditions. A single success is not grounds for meeting the rating. Limits arising from factors that simulation does not reproduce, such as heat, current and wear, are indicated as items that can be confirmed only by physical testing (**PAVS-ITM-10**).

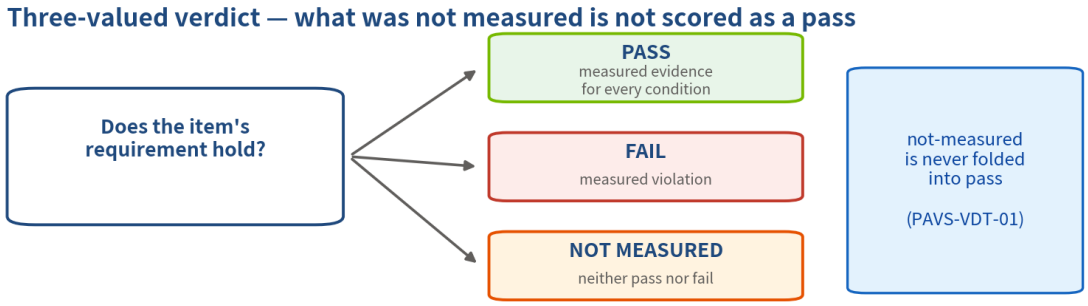
4.2 Measurement — distinguishing the existence of a value from a measurement

Every value used in scoring shall be accompanied by **proof of measurement** (provenance).

- The source of a value (measured/commanded value/default value/estimate) is declared, and a value that is not measured shall not be used in scoring an item requiring measured values.
- Missing is not 0. If there is no sample, the value shall be "none", and a missing value shall not be filled with 0 or a default value — a structure in which a stopping time of 0 seconds is recorded and a full score is produced even though no emergency stop was performed is prohibited.
- A measurement channel shall be confirmed to actually respond to signals (**liveness**). If the measurement value stays the same while the input state changes, that value is not a valid signal but an indication of a measurement failure.
- Whether a measurement channel is sound is not confirmed by reading back values configured inside the channel (a self-consistency cross-check). It is confirmed by an independent check that **injects a reference stimulus of known magnitude and observes whether that stimulus reaches the very last output of the measurement path (the final record)**.
- For a proxy metric that does not measure a physical quantity directly and whose value varies with computational settings (for example, a contact force figure inversely proportional to the simulation time step), the range and conditions within which it can be compared shall be stated, and it shall not be used to compare absolute values between results obtained under different conditions.

> **EXAMPLE** — Not asking for the source of values goes wrong in both directions. In one actually measured case, a maximum-speed metric was in fact the driving command value multiplied by a coefficient: the maximum speed of a robot that actually ran at 1.4 m/s was recorded on paper as 4.7 m/s (measured 2026-08), a state in which a false fail of a safety clause could be produced, and the nonexistent overspeed actually turned on the violation counter. In the opposite direction, a collision count of "0" that nobody had measured let a safety item pass (false pass). Declaring the source of values blocks these errors in both directions with the same mechanism.

5. Link 3 — verdict: not-measured is a third state



> **What was not measured is not folded into a pass — not-measured is a third state that is neither pass nor fail.**

The core of the verdict rules is one thing: what was not measured is not folded into a pass.

- A verdict shall be not 2-valued, pass/fail, but 3-valued: **pass/fail/not-measured**. This 3-valued verdict inherits the pass/fail/inconclusive of ISO/IEC 9646. Not-measured is not counted as a pass, and is folded into neither the denominator nor the numerator of either the pass rate or the fail rate. However, where a mandatory item (an item for which a pass

cannot be established without its measurement) lacks measurement, it may be treated as a fail for the reason ‘absence of measurement’. Even in that case, the count of not-measured results is retained and indicated as a separate state distinguished from fails for absence of measurement.

- A pass shall be produced only from **positive evidence that all required conditions were met as measured**. Neither a default value of the verdict logic nor a path for unmet conditions shall return a pass.
- A success metric shall be **cross-validated against independent evidence of the physical event** that it presupposes. A structure in which success is counted even though the robot's execution mechanism did not actually move (no-op) is prohibited.
- Where there are multiple stages that give verdicts, the operator shall state the strictness order of those stages in documentation. A more lenient stage shall not overwrite a fail or warning given by a stricter stage with a pass.
- A measurement failure is not filled in with substitute values or mock values but is exposed and reported as a failure (fail-loud). In a test declared to be scored by physics simulation, using as evidence results produced through a path that bypasses physics computation (for example, a kinematic path that moves things by setting positions directly) in order to avoid a defect in the physics computation path is prohibited.
- The scorer shall not embed the performance limits of the SUT as constants in its code. All limits shall derive from the declared specifications of the SUT (a machine-readable manifest) — this is a prohibition of structures in which the scorer distorts the SUT.
- **When a defect in a verdict criterion is corrected, all existing verdicts from the defect period shall be retrospectively audited against the new criterion and the results recorded.**

> **EXAMPLE** — The incident that gave rise to the last clause: in one scoring system, the default path of the success verdict was returning True even when conditions were not met. In the retrospective audit after the correction, a census of all 401 passes in the defect period found 89% to be false passes (retrospective audit 2026-05). Had there been no obligation of retrospective audit, this figure would not exist, and the false passes would have remained on record as achievements.

6. Link 4 — verification: scoring the scorer

> **Trust in a scoring system comes not from claims but from the tests that the scorer itself has passed.**

The requirements for scorer verification are as follows.

- **Permanent negative controls**: The scorer passes only if a scenario that deliberately violates an item fails on exactly that item. Automatic comparison against the expected verdict (expect) is kept resident in the regression suite.
- **Qualification testing**: When the scorer is changed, it shall be proven that re-scoring a set of known correct answers (golden set) agrees with the existing verdicts (verdict shifts are counted and disclosed).
- **Confirmation of effective values in comparative tests**: The independent variable of an A/B test shall be confirmed in each test arm by the effective value actually applied by the system, not by the intent of the setting. A test whose effective values are unconfirmed is not a refutation; it is unverified.
- **Checking the measuring instruments before blaming the SUT**: Before concluding that a test failure is a defect of the SUT, whether the measurement system is sound shall first be confirmed by a refutation test. Measurement defects easily look like SUT defects.

- **Confirming the authenticity of model specifications (model authenticity)**: The simulation model of the SUT (technical specifications, geometry and material properties) shall be verified by comparison against the specifications confirmed by the manufacturer. Where the manufacturer publishes multiple values for the same entry, the value adopted as the reference and the reason for it are stated, and for each entry in which the model value differs from the reference, the grounds for the difference are recorded. A difference whose grounds are not recorded is treated as unverified, and the results of a model in which a mismatch is confirmed in the comparison cannot be used as grounds for a conformance claim (**PAVS-VRF-05**). It has been confirmed by measurement that, for the same robot and the same task, the publicly distributed technical specifications and the manufacturer-confirmed technical specifications produce opposite verdicts (a false pass of a pose impossible on the physical robot ↔ refusal to execute) (measured 2026-08). In a system that scores third parties, the authenticity of model specifications is a precondition of scoring.
- **Recording failed methods**: Failed verification and measurement methods and the evidence that they were wrong are entered in an official register (the dead-end registry), and a step of consulting that register before starting a new investigation is made part of the procedure. When a wrong diagnosis is corrected, the original record is not deleted but is preserved, linked to the correction record. The history of misdiagnoses also becomes evidential material for the next verification.
- **Reproducibility**: Every verdict shall be reproducible by recomputation with a single command, and the seed, configuration and scoring code version shall be recorded with the result in fixed form. When the scoring code changes, whether the existing verdicts produced by that code remain valid shall be re-evaluated.
- **Standing execution of self-checks**: The self-checking mechanisms of a scoring system, such as negative controls and injection checks, shall be executed automatically at defined intervals or whenever the scoring system changes. The result of a check that was not executed shall not be reported as ‘no violation’, and where automatic execution does not cover all of the protected verdict paths, the omitted scope is stated (**PAVS-VRF-16**).

> **EXAMPLE** — Defects in model specifications pass without errors. In one publicly distributed robot model, gravity was disabled on all of the arm links. The simulation ran without errors, and the precision measured with that model was a value that excluded the weight of the arm itself. A check confirming the gravity state of the links bearing the test load (a fail-loud check that exposes the run as a failure if there is an anomaly) revealed this. The remedy was carried out in the order of model normalization, re-measurement and re-judgment, and recording and pre-checking of the per-link gravity state (2026-10, Casebook D12).

> **EXAMPLE** — How the self-check clause came about: the previous edition of the specification required only that self-checking mechanisms ‘be provided’. In practice, there was a checker that was implemented but not connected to an automatic execution point. A self-check on 2026-08-26 revealed this; **PAVS-VRF-16** was newly established to fill that gap, and it was counted as not met by the publisher itself.

> **NOTE** — As grounds for trusting simulation results, the automated driving regime required "proof of the suitability of the tool" (R157). PAVS follows the same structure. Correspondence (agreement) between simulation and physical hardware is not a claim but an **object of measurement** (an academic precedent for quantifying the correlation: SIMPLER 2024). Calibration gives priority to finding and adding the physical phenomena missing from the model (missing physics — for example, joint friction) over fitting values by increasing the number of parameters. A calibration value is adopted as the official value only when it also fits under test conditions that were set aside and not used for calibration (held-out tests).

7. Link 5 — reporting: performance boundaries, not scores

> **The standard for results reporting is not ‘how many points’ but having the document itself describe ‘the range within which this score can be trusted’.**

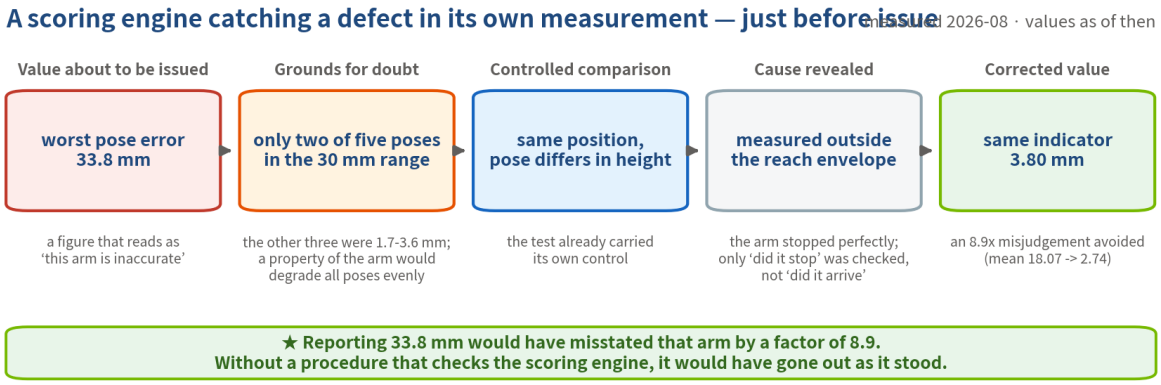
The requirements for results reporting are as follows.

- **Attribution disclosure**: A score report shall indicate, item by item, whether the grounds for each item's verdict come from an authoritative reference standard or are the issuer's own definition. Issuing a score report under a standard name not registered in the issuing system, that is, a nonexistent standard name, is prohibited.
- **Coverage disclosure**: What was measured and what could not be measured are stated as a **list**, not as a ratio.
- **Limits of self-declaration**: A verdict shall not be raised by success stated by the scenario itself (self-declaration). A verdict produced by self-declaration alone cannot be grounds for a results document or a conformance indication (blocking circular reasoning). Self-declaration in the direction of lowering a verdict (where a scenario itself states a fail) is permitted, and where self-declaration has contributed to the overall verdict, that fact is indicated item by item.
- **Boundary notes**: The "meaning" of a pass (whether it is a runtime guarantee, a capability boundary, or untested because there was no stimulus) shall be included item by item. It is recommended to provide a mechanism by which a measurement contradicting that note is automatically flagged (**PAVS-REP-14**).
- **Statistical discipline**: Success-rate-type figures are reported only together with the number of trials and confidence intervals. An improvement claim is established only when the null control results are reported with it. A null control is a comparison condition that passes through the same processing path unchanged with only the claimed improvement effect removed, and it allows the difference produced by the processing path itself to be distinguished from the actual improvement effect.
- **Statement of limitations**: Known simplifications, approximations and dormant defects are disclosed with their range of validity and direction of distortion (optimistic/pessimistic) stated.
- **Indication of the physical-hardware verification stage (★core)**: A document reporting simulation results shall indicate how far the simulation model has been compared with physical hardware — ① only self-consistency confirmed ② confirmed up to cross-comparison with other computation methods or models ③ compared with physical-hardware measurement under the same conditions, with the difference within the declared range. If there is no correlation evidence, ‘none’ is written, and the indication itself is not omitted (**PAVS-REP-15**).
- **Discipline of completion declarations**: When a report claims ‘complete’, it shall also state, in a defined format and for each promised deliverable, whether it was achieved, deferred or missing, and partial achievement shall not be indicated as complete.
- **Integrity**: Issued score reports shall be managed in a ledger that can only be appended to (append-only), so that tampering with, deletion or reordering of records is mechanically detectable.
- **Declaration of conflicts of interest**: Where the issuer of the scoring has been involved in the adaptation or training of the SUT, that fact and the scope of independence are always declared in the score report. "Neutral" can be indicated only for engagements in which there was no involvement in adaptation.

> **EXAMPLE** — There is a measured case in which two paths issued opposite score reports for the

> same test run: the issuing tool equipped with honesty mechanisms returned a fail, while another screen path returned "Pass — certified for operation". The latter generated the same score report even from empty input — because it substituted 0 for absence, fabricated default values, and let the requester choose the standard name. The clauses on attribution disclosure, blocking of nonexistent standard names and coverage disclosure came from this incident.

8. Origin of the clauses — the place of the Casebook (Part 2)



One case (2026-08) — how the scorer caught a defect in its own measurement just before issue. The clauses of this specification came from incidents like this.

> **The clauses of this specification came from errors that actually occurred. Every clause is mapped to its grounds of origin — an incident from in-house operations, a failure record in the published literature, or the acceptance conditions of a precedent.**

Just as aviation safety regulations came from accident investigations, the clauses of PAVS were also generalized from recorded events. Every clause is mapped to its grounds of origin — an incident from in-house operations, a failure record in the published literature, or the acceptance conditions of a precedent (the correspondence table is in Part 1 Annex B, and the incident narratives are in Part 2). Among these, an incident from in-house operations is **an actual event in which, in the absence of that clause, a wrong result was produced without any error indication**.

The Part 2 Casebook is that record, and each case is described in the following 4-step structure.

> **Problem situation → type of error in the absence of the specification (false pass/false fail) → the mechanism that detects it → method of reproduction**

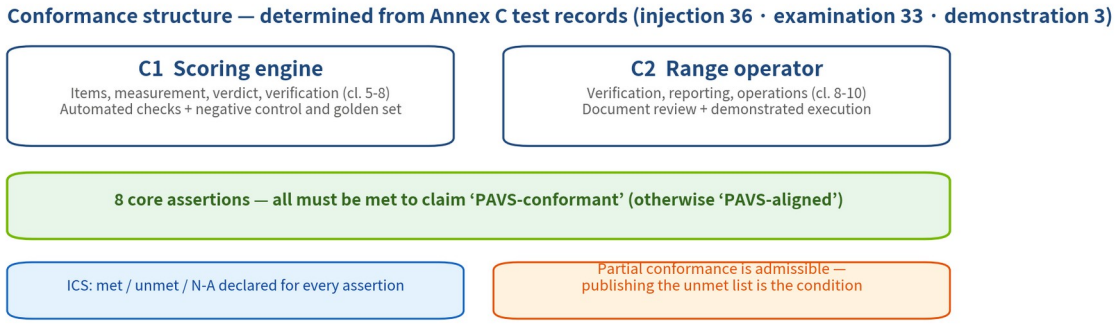
The cases are not a list of achievements. Many of the cases are records in which "the verification mechanisms of the scoring system caught defects of the scoring system itself".

A structure that applies the same criteria to itself is the target state required by this specification. The representative case groups are as follows.

- Source of values (mistaking commanded values for measurements → false fail; absence as 0 → false pass)

- Confusion of units and meaning (a velocity value passing a force item — even where the clause number of a reference standard is the same, it can denote a different physical quantity depending on the edition)
- Mismatch between an item's name and what is actually measured (a separation distance item was in fact measuring distance traveled)
- False passes created by verdict default values, and retrospective audit (89% of the passes in the defect period were false passes — a census of all 401 passes · retrospective audit 2026-05)
- Negative controls of the scorer (tests that confirm whether the scorer can produce a fail)
- Authenticity of model specifications (opposite verdicts between publicly distributed technical specifications and manufacturer-confirmed technical specifications; a public model with gravity disabled on the arm links)
- Liveness of measurement (a case in which control that trusted a faulty measurement value performed worse than control without measurement)
- Verification of the operation of safety mechanisms (not whether a mechanism is invoked but whether it actually operates when invoked — injection tests and placebo controls)
- Non-execution of self-checks (the checker was implemented but not connected to an automatic execution point)
- Turning third-party safety into items (if the test baseline has no third-party conditions, the same robot on the same path can both pass and fail)

9. Conformance — requirements that come with test methods



> **Every requirement is assigned an assertion identifier, and a test method is specified for all 72 — injection 36, audit 33, demonstration 3. Conformance is determined from the records of tests performed by these procedures, not from document review alone.**

9.1 A test method for every requirement

PAVS extends the ISO principle — "requirements shall be objectively verifiable, and only verifiable requirements are included in a standard" — one step further.

Every requirement of Part 1 is assigned a test assertion identifier, and a test method is specified for all 72 of them — injection 36, audit 33, demonstration 3. Conformance is determined from the records of tests performed by these procedures, not from document review alone. This is the design philosophy that "the substance of the specification is the corpus", applied to the determination of conformance.

Each test method (normative Annex C) contains the preconditions, the procedure, the pass criteria, the record items and **the routes by which the test itself can be passed**

wrongly (pitfalls). Assertions designated for automated checking are, in principle, tested by **injection tests** (tests that deliberately insert a defect and observe whether it is detected). This is because the fact that a checker is installed is not evidence that the checker reacts to actual defects. Assertions that concern the properties of records over a period are confirmed by audit tests, and assertions for which the defect condition cannot be produced, or cannot safely be produced, are confirmed by demonstration tests.

> **A requirement without a test method is not reproducible — if two testers each test the same system and obtain different results, the specification cannot serve as a basis for verdicts.**

9.2 Conformance classes

Conformance is defined by role (W3C conformance convention). Interpretation guidance for consumers who read the scores is provided not as a normative conformance class but as informative guidance (this White Paper and the Casebook).

Class	Target	Meaning of conformance
C1 Scorer implementation (scoring engine)	Simulation scoring system	Requirements of Clauses 5–8 of Part 1 (items · measurement · verdict · verification) met
C2 Range operator	Organization that conducts tests and issues results	Requirements of Clauses 8–10 of Part 1 (verification · reporting · operations) met

The form of a conformance claim is given in Part 1 11.2, and the denominator of the Implementation Conformance Statement (ICS) is Annex A (72 assertions). Domain profiles (domain-specific selections of items, such as manipulation, mobility and safety) are declared in separate public documents (**PAVS-CNF-05**).

9.3 Reference implementation

PAVS has a reference implementation — an operational system comprising scenario domain-language execution, measurement-provenance wiring, a 3-valued verdict processor, a negative-control suite and an append-only score ledger. Part 3 (Reference Implementation & Reproduction — published) records its mechanisms and 3 demonstrations. Demonstration A is CPU-only, runs with a single command, needs no GPU and automatically compares its output against reference values. The reproduction package for Demonstration A is provided on request (execution commands and internal paths are provided to joint pilot-test partners). That the three demonstrations differ in their level of disclosure is set out in a table in Part 3 §5.

The goal of the reference implementation is to prove that every clause is executable, and its current status of assertions met (test of 2026-09-04 · 7 of 72 met) is disclosed in full in the Conformance Statement. The reference implementation also serves as a reference point that adopters can cross-check in their own environments.

Because the body of the specification is neutral with respect to particular products and organizations, Part 3 treats the implementation anonymously. The configuration, boundaries and form of delivery that the implementation has as a product are described, under its name, in **Part 4 Solution Profile** (Part 4 is unpublished — to be published next). Other implementations can write their own profiles in the same format. The actual form of a conformance claim that states its assertions not met is shown by the Conformance Statement (9.4).

9.4 The party that wrote the specification is tested first

The publisher applied this specification to its own reference implementation first. The test was performed on 2026-09-04 by the procedures of Annex C (target specification Part 1 v1.2; Class C1 (scorer implementation); REP-15 retested on 2026-09-06; CNF-01 and CNF-02 retested on 2026-10-20; v1.3 and v1.4 are editorial revisions with the same assertion requirements).

- Met: 7 of 72 (MEA-02★, VDT-01★, VDT-06, VRF-02, VRF-03, REP-04 and REP-15★) · 3 of the 8 ★core assertions
- ★Core not met: MEA-01, VDT-02, CNF-01 and CNF-02 (CNF-01 and CNF-02 = retest of 2026-10-20 — the class and the lists were omitted from summary documents that restate only the met count)
- ★Core inconclusive: VRF-01 (a state in which being met could not be confirmed from the test records)
- Conclusion: because not all ★core assertions are met, conformance is not claimed and only “PAVS-aligned” is indicated
- Disclosure: all 65 assertions not met are listed in Appendix E and in the Conformance Statement (first published 2026-09-09 · current edition = document list at the publication location)

This result is the publisher's self-test, not a third-party audit.

Self-application also changed the specification itself. The record consists of the following 3 entries.

- **PAVS-VRF-16 newly established (2026-08-26)** — a self-check revealed a checker not connected to an automatic execution point
- **PAVS-REP-15 promoted to ★core (2026-09-01)** — the publisher itself had been issuing score documents without sim-to-real correlation evidence
- **Correction of the wording of PAVS-VDT-01 and REP-04 (2026-09-04)** — the self-test revealed a mismatch between the wording of the specification and its intent. Because aligning the implementation with the wording would have made verdicts more lenient, the wording was corrected, and notice of it was given as a change in requirement content

10. Position in the ecosystem

> **PAVS is a layer, not a replacement. Where existing standards specify what is measured, PAVS specifies how measurement, trust and reporting are done in simulation.**

PAVS is not a replacement but a **layer**.

- **With existing standards**: ISO 9283, ISO 18646, ISO 3691-4 and ISO 10218 are the authoritative source for "what to measure". PAVS **extends and complements** those measurement items at the layer of "how they are reproduced, measured, scored and reported in simulation". Each item links a reference standard number as its grounding standard (anchoring), and for application in Korea the harmonized KS standards (KS B ISO 10218, KS B ISO 3691-4, etc.) are stated alongside — because there is a time lag between the international editions and the KS editions, dual indication is the accurate form.
- **With physics engines and simulators**: It is not a specification of any particular engine. The reality that verdict differences between engines are actually measured is the basis for the clauses on cross-engine comparison and on verifying the authenticity of

model specifications. A simulator is a tool for the stage of developing a robot, and this specification addresses the proof required at the stage of shipping and deploying a robot — the two do not compete but connect as successive stages.

- **With world models and synthetic data**: Generative augmentation is not a replacement for scoring but an auxiliary layer, and the statistical discipline of this specification is that whether to adopt it shall also be decided by measured comparison (accompanied by a null control).
- **With national infrastructure**: Virtual field-testing infrastructure (test fields) and scoring methodology are complementary. What is scored on top of the infrastructure, and how it is scored, is the area this specification addresses.
- **With statutory certification**: PAVS is not a certification scheme. Its target state is to be used at the layer of pre-verification and regression verification for statutory certification (for example, the Outdoor Mobile Robot Operational Safety Certification), and it does not appropriate the vocabulary of accreditation.

11. Open specification roadmap

> **The standardization path has 3 stages: open specification + reference implementation → consortium → formal standardization. This suite corresponds to stage 1 of that path.**

Three stages on the path to becoming a standard have been proven in practice.

1. Open specification + reference implementation	2. Consortium	3. Formal standardization
Publication of the specification · proof of 'every clause executable' through the reference implementation (goal) · securing adoption cases (de facto standard)	A consortium adjacent to scenario languages (ASAM lineage) or an independent consortium — transition via the lightweight deliverable route (PAS · TS · IWA)	Securing a number as an ISO lightweight deliverable, then promotion to IS, or a new work item proposal to ISO/TC 299 — a relationship of extension and complementation with ISO 18646

The industrial 3D scene description standard (OpenUSD family) is following the same sequence — de facto standardization of an open implementation → formation of a consortium → pursuit of formal standardization.

Stage 1 of this path is this suite. It publishes **a readable specification (Parts 0 and 1), a casebook with evidence (Part 2) and an executable reference (Part 3)**. The named Solution Profile (Part 4) has not yet been published and will be published next.

References

These are the external sources cited in the body. The verification date is based on the primary-source check of 2026-08-25 (partially rechecked 2026-08-27), and whether Korean certification criteria contain clauses recognizing simulation is based on the survey of 2026-09.

International regulations and assessment protocols

- UNECE, UN Regulation No. 157 — type-approval regulation for Automated Lane Keeping Systems (ALKS) (2021)
- UNECE WP.29, New Assessment/Test Method for Automated Driving (NATM) (adopted 2021-06)

- Euro NCAP, press release on the 2026 protocol changes (2025-11-26) · virtual testing protocol (published 2025-03)
- ASAM, OpenSCENARIO · OpenDRIVE
- Regulation (EU) 2023/1230 on machinery, EUR-Lex 32023R1230 (Article 54 · Recital 55)

Standards and standards documents

- ISO 9283:1998 · ISO 10218-1/-2:2025 · ISO 3691-4:2023 · ISO 18646 series · ISO 13482 · ISO 25785-1 (development initiated)
- KS B ISO 9283 · KS B ISO 10218 · KS B ISO 3691-4:2021 · KS B ISO 13482
- ISO/IEC 9646 family · ISO/IEC TS 42119-2:2025 · ISO/IEC 17025 · ASME V&V 40 · NASA-STD-7009
- ISO/IEC Directives Part 2 · W3C QA Framework
- IEEE humanoid standards study group report (2025-09)

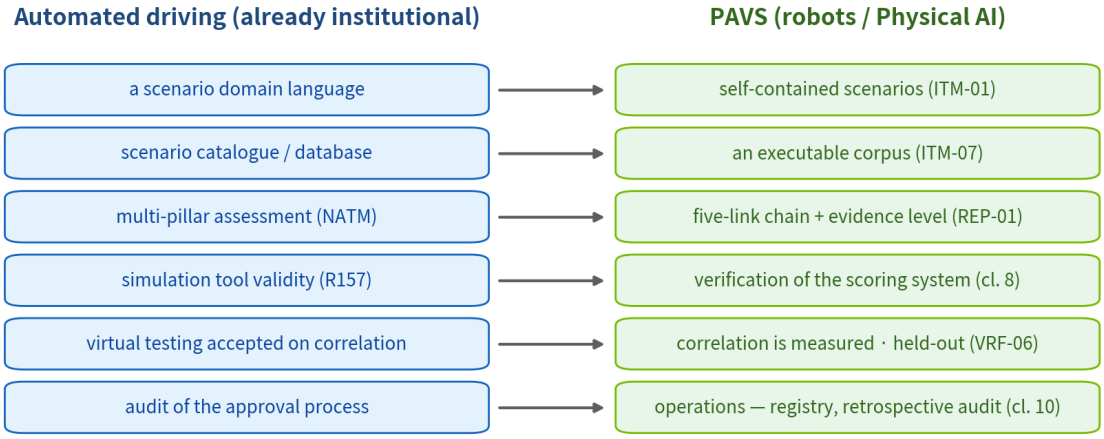
Korean legislation and policy

- “Framework Act on the Development of Artificial Intelligence and the Establishment of a Foundation for Trust” (Korea) (인공지능 발전과 신뢰 기반 조성 등에 관한 기본법) (in force 2026-01-22)
- “Intelligent Robots Development and Distribution Promotion Act”, Article 40-2, Outdoor Mobile Robot Operational Safety Certification (in force 2023-11-17) · review items published by the Korea Institute for Robot Industry Advancement (KIRIA)
- Ministry of Science and ICT, “Strategy for Securing Core Competitiveness in Physical AI” (피지컬 AI 핵심 경쟁력 확보 전략) (2026-07-01)
- National Robot Test Field (Korea), field-testing support platform · technology development of a digital-twinning framework (2024–2028)

Academic literature

- LIBERO-PRO, arXiv 2510.03827 (2025)
- LIBERO-Plus, arXiv 2510.13626 (2025)
- Snyder et al., number of trials and statistical testing in comparing robot manipulation policies, arXiv 2503.10966 (2025)
- Cross-simulator benchmark of deformable-object (cloth) manipulation, IEEE Robotics and Automation Letters (2024), arXiv 2310.09543
- Humanoid-Gym, arXiv 2404.05695
- Toyota Research Institute, evaluation study of large behavior models (LBMs), arXiv 2507.05331 (2025)
- SIMPLER, arXiv 2405.05941 (CoRL 2024)

Appendix A (informative) — Correspondence between automated-driving precedents and this specification



the correspondence is structural, not an equality of institutional standing

This is the correspondence between each element of the automated-driving verification ecosystem and this specification. The correspondence is a **similarity of structure**, not an identity of institutional status — PAVS is an open specification at the proposal stage.

Element of the automated-driving ecosystem	Problem solved	PAVS correspondence
Scenario domain language (ASAM OpenSCENARIO/OpenDRIVE)	Test scenarios in a language that people read and machines execute — a lineage in which regulatory text becomes executable code	Requirement for self-contained domain-language scenarios (**PAVS-ITM-01**)
Scenario catalogs · databases (lineage of national preliminary research → incorporation into regulation)	The substance of a test as a set of scenarios, not a document	Executable corpus — scenarios + controls + expected verdicts (**PAVS-ITM-07**)
Multi-pillar verification (UNECE NATM: simulation · track · real road · audit · in-service monitoring)	Verdicts that combine evidence from different layers	5-link chain of trust (item–measurement–verdict–evidence–reporting) + evidence levels L0–L3 (**PAVS-REP-01**)
Proof of the scope and validity of simulation tools, and toolchain verification (UNECE R157)	Institutionalization of the grounds for trusting tools	Requirements for verification of the scoring system (Part 1 Clause 8) — negative controls · golden set · verification of the authenticity of model specifications · confirmation of measurement liveness
Acceptance condition for virtual testing: proof of correlation with real-world measurement (Euro NCAP 2026)	Qualification of simulation evidence by comparison with real-world measurement	Agreement claims accompanied by measurement method · sample · uncertainty · formal adoption only of values that pass a separate condition not used for calibration (held-out test) (**PAVS-VRF-06**) · indication of the physical-hardware verification stage (**PAVS-REP-15**)
Process audit in type approval	Evidence of procedure, not only of results	Operational requirements (Part 1 Clause 10) — register of failed

		methods · retrospective audit · record-keeping discipline

Appendix B (informative) — Correspondence table of reference standards and KS

These are the reference standards to which the items of each domain profile are linked as grounding standards (anchoring), and their status of harmonization in Korea. Verified as of 2026-08, based on publicly available material. **This specification does not imply conformance certification to any standard in the table.** Because there is a time lag between the international editions and the KS editions, citation follows, as a principle, dual indication: direct reference to the international edition + the KS stated alongside.

Domain (profile candidate)	International reference standard	KS harmonization (based on public search)	Layer addressed by PAVS
Manipulator performance	ISO 9283:1998	KS B ISO 9283 exists	Simulation reproduction · scoring · reporting of measurement items such as pose accuracy · pose repeatability
Industrial robot safety	ISO 10218-1/-2:2025 (collaborative requirements integrated)	KS based on the 2011 edition — harmonization of the 2025 edition not confirmed (unknown)	Linking of safety-clause items to grounding standards · dual indication for the edition time lag
AMRs · driverless industrial trucks	ISO 3691-4:2023	KS B ISO 3691-4:2021 (harmonization basis = ISO/DIS 3691-4.2, IDT — KS harmonizing the 2023 edition not confirmed)	Linking of safety items such as zone speed · stopping to grounding standards
Service robot performance	ISO 18646-1–4 (new parts under development)	Harmonization status unknown	Linking of locomotion · navigation · manipulation items to grounding standards — relationship of extension and complementation
Personal care robot safety	ISO 13482	KS B ISO 13482 exists	Linking of service-environment safety items to grounding standards
Korean institutional interface	—	Outdoor Mobile Robot Operational Safety Certification under the Intelligent Robots Act (Article 40-2, in force 2023-11-17 — 8 review items based on physical hardware, published by KIRIA · no clause recognizing simulation results	Pre-verification · regression verification layer for statutory certification (no false claim of certification)

		found in the certification criteria, 2026-09 survey)	
Industrial safety certification scheme	—	**Unconfirmed** — whether industrial robots · collaborative robots are covered by the safety certification scheme of the Occupational Safety and Health Act (Korea) (Ministry of Employment and Labor · Korea Occupational Safety and Health Agency), and the names of the applicable items, could not be confirmed from primary sources	If confirmed, to be linked as grounding standards for safety items
Korea Certification (KC)	—	**Unconfirmed** — the applicable scope of robot products (product groups · applicable laws) within the body of laws subject to the KC mark could not be confirmed from primary sources	If confirmed, to be linked as grounding standards for product safety items

> **NOTE — Indication of "unconfirmed".** The two rows above are not empty because they were not surveyed; **they are indicated as unconfirmed because the primary-source check could not be completed**. The measurement principle of this specification (not-measured is not folded into measured) is applied to the specification's own statements as well. They will be reflected in a revised edition as soon as they are confirmed.

B.1 Position on interfaces with Korean standards

This clause sets out the position of this specification on its relationship with Korean standards and certification schemes.

First, because the layers differ, they do not compete. Korean and international standards set **what is measured and the pass criteria**, and this specification sets **how those measurement objects are reproduced, measured, judged and reported in simulation**. This specification **does not specify pass thresholds** (Clause 1, Scope), and the source of a threshold is exposed as it is in the score report (**PAVS-REP-02**). A division of labor holds in which Korean standards supply the thresholds and this specification governs the measurement and reporting against those thresholds.

Second, Korean mapping is work at the profile layer. This specification already specifies the domain profile mechanism — a profile selects a set of items but shall not relax the requirements (**PAVS-CNF-03**), and shall declare its name, item list, selection rationale and excluded items in a public document (**PAVS-CNF-05**). A set of items that links Korean standards as grounding standards is accommodated **not by revising the body of the specification but by establishing a new profile**. The body of the specification remains nationally neutral.

Third, the current status is described as it is. Based on the table above, KS harmonization is partly confirmed and partly unconfirmed, and the scope of application of the industrial safety certification scheme and of KC is unconfirmed. Because there is a time lag between the international editions and the KS editions, citations **state the international standard number and the KS number side by side**.

Fourth, Korean mapping is not a matter to be carried out unilaterally. Which items of which Korean standards are adopted as items is a matter for agreement among the competent bodies and stakeholders, and if the authors of this specification decided it alone, the selection itself would become a product of their interests. The selection of items for a Korean profile is therefore **proposed to be carried out jointly with the competent bodies**, and this specification provides the methodological layer on which that joint work can stand.

> **NOTE** — The statements in this clause do not mean that this specification conforms to Korean standards or has been certified in accordance with them. Citation of standard numbers is linking to grounding standards, and the vocabulary of accreditation is not used.

Appendix C (informative) — Wording boundaries for citation

These are wording boundaries for intermediaries who cite this document for translation, presentation, sales or design purposes. Even if the source material stays within the boundaries, a quotation that crosses them produces the same incident. When the text of the suite is revised, this table is revised with it.

What is to be said	Render it like this (○)	Not like this (×)
Status of the specification	"open specification at the proposal stage"	"international standard" · "KS/ISO certified" (implies approval by a standards body)
The word "standard"	"standard items and standard scenarios defined by this specification"	"standard certification scenario" (implies certification by an external body)
Reference standards	"links the measurement items of ISO ○○ as a grounding standard (reference)"	"has obtained ISO ○○ conformance or certification"
Reference implementation	"the goal of the reference implementation is to prove that every clause is executable, and its current status of assertions met (7 of 72 met) is disclosed in full in the Conformance Statement"	"the reference implementation executes all assertions" (present-tense statement of fact) · statement of fact that it is a finished product or a service on sale
Reproduction package	"the reproduction package for Demonstration A is provided on request"	"anyone can download and reproduce it" (states public availability as fact while it is not posted)
Score reporting	"specifies the score reporting format and integrity mechanisms"	"we issue score reports for you" (present-tense statement of a service as fact)
Neutrality	"conflicts of interest are mandatorily declared in the score report" · "the publisher is not a certification or	"neutral third-party body" (claim of status in the present tense)

	accreditation body"	
Case figures	The physical quantities of the incident as they are + measurement date stated alongside	Upgrading by rounding · turning a possibility into a statement of fact ("could have been produced" → "received the verdict")
Failure cases	"the system caught and fixed its own defect" (3 steps: discovery → correction → re-verification)	Dramatizing the defect in isolation · omitting the fact of correction
External research	"the authors reported/claimed"	Stating external claims as fact
Conformance	"partial conformance presupposing that the core assertions are met — list of assertions not met disclosed"	"PAVS passed · certified" (implies that assertions not met are not disclosed)
Status of the publisher	"self-test result 'PAVS-aligned' — all assertions not met disclosed"	"PAVS-conformant implementation" · wording as if it were the result of a third-party audit

Appendix D (informative) — Glossary (plain-language explanations)

The normative definitions in Part 1 Clause 3 are authoritative; this glossary is a plain-language explanation to aid understanding.

- **System under test (SUT)** — The robot or AI that is the object of a test.
- **Item / test assertion** — An individual unit of scoring / a single statement, making up that item, whose truth or falsity can be determined. Each clause of this specification has an assertion number (PAVS-XXX-NN), and a test method (injection, audit or demonstration) is specified for every one of the 72.
- **Measurement provenance** — An entry indicating whether a figure is a measured value, a commanded value, a default value or an estimate. A value whose provenance is not declared shall not be used in scoring.
- **Missing** — The state in which no measurement was performed. It means the absence of a value, not 0.
- **Absence of measurement** — The state in which no measurement was made for a mandatory item (an item for which a pass cannot be established without its measurement). A fail may be given for this reason, in which case the reason for the fail shall be identifiable in the record.
- **3-valued verdict** — Pass/fail/not-measured. An item that was not measured is not treated as a pass. An absence of measurement of a mandatory item may be treated as a fail with its reason stated, and even then the count of not-measured results is retained and indicated separately. It is a concept inherited from the pass/fail/inconclusive of ISO/IEC 9646.
- **fail-loud** — The principle of explicitly reporting a measurement failure as a failure, without compensating for it with substitute values.
- **Negative control** — A test target constructed to commit a deliberate violation. If the scorer passes it, the scorer fails.
- **Null control** — A comparison condition that goes through exactly the same processing path with only the claimed improvement effect removed. It distinguishes the difference produced by the processing path itself from the actual improvement effect.

- **Placebo control** — A test confirming that an alarm does not go off under conditions in which it is not supposed to go off.
- **Golden set** — A reference set of tests whose verdicts are settled. After the scorer is changed, it is re-scored to confirm whether the verdicts remain unchanged.
- **Effective value** — The value the system actually applied. Configuring a value does not mean it was applied.
- **Boundary note** — An annotation describing, item by item, the scope that a given pass guarantees.
- **Evidence level (L0–L3)** — Code exists (L0) → unit test (L1) → end-to-end execution on the production path (L2) → direct inspection of artifacts (L3). A pass is issued only at L2 or above.
- **Model authenticity (verification of the authenticity of model specifications)** — Confirming whether the technical specifications, geometry and material properties of a simulation model match the manufacturer's confirmed specifications and, where they differ, whether there are recorded grounds for the difference.
- **Held-out test** — A test condition set aside and not used for calibration. A calibration value is formally adopted only when it also holds under this condition.
- **Retrospective audit** — The procedure of, after a defect in the verdict criteria has been fixed, re-judging all existing verdicts from the period in which the defect existed against the new criteria, and recording the direction and magnitude of the verdicts that changed.
- **Dead-end registry** — The official register of approaches that have already failed. It is a mechanism to prevent the same failure from being repeated, and even corrections of misdiagnoses are preserved linked to the original record, without deleting it.
- **Manifest** — A declaration file containing the identification information and specifications (joint and actuation limits) of the SUT. The scorer rejects directives outside this range instead of approximating them.
- **Profile** — A domain-specific selection of items (manipulation, mobility, safety, etc.). It may choose items but shall not relax requirements.
- **Score report** — A document containing test results, issued by a range operator. It does not mean a statutory test report.
- **Conformance Statement** — A document stating, in a prescribed form, whether an implementation meets each assertion of the specification. It is valid only if it attaches the checklist (ICS) of all assertions of Annex A and discloses every assertion not met (Part 1 11.2 and 11.3).
- **PAVS-aligned** — The distinguishing expression that an implementation with even one ★core assertion not met uses instead of 'PAVS-conformant' (Part 1 11.2).
- **★Core assertion** — The minimum set presupposed by a 'PAVS-conformant' claim (**8** — 8 since the promotion of REP-15 on 2026-09-01). The authoritative list is Part 1 11.2.1. If even one is not met, only "PAVS-aligned" can be used.

★Core assertion	Meaning
PAVS-MEA-01	Provenance (measured · commanded value · default value · estimate) stated for each value used in scoring, and items requiring measurement scored with measured values only
PAVS-MEA-02	No filling with 0 or default values when there is no measurement
PAVS-VDT-01	3-valued verdict of pass · fail · not-measured · not-measured not folded into pass
PAVS-VDT-02	A pass produced only from evidence that all required conditions were met by measurement
PAVS-VRF-01	Permanent maintenance of deliberately violating scenarios (negative controls) and automatic

	comparison with expected verdicts
PAVS-REP-15	Indication of the stage to which the simulation model has been compared with physical hardware and of the status of correlation evidence · ‘none’ if there is no evidence
PAVS-CNF-01	Conformance claimed in the prescribed form, with every assertion not met disclosed
PAVS-CNF-02	Attachment of a checklist (ICS) of whether each Annex A assertion is met · restriction on the eligibility of not applicable (N/A)

Appendix E (informative) — Self-test checklist (Annex A, 72 assertions)

These are the per-assertion verdicts of the 9.4 self-test. Target specification Part 1 v1.2 · Class C1 (scorer implementation) · tested 2026-09-04 (REP-15 retested 2026-09-06 · CNF-01 and CNF-02 retested 2026-10-20). ★ marks a core assertion. Not applicable: 0. The test records and invalidity-condition reports for each verdict are carried in the Conformance Statement.

Clause group	Met	Not met
ITM (0/10)	—	ITM-01 · ITM-02 · ITM-03 · ITM-04 · ITM-05 · ITM-06 · ITM-07 · ITM-08 · ITM-09 · ITM-10
MEA (1/8)	MEA-02★	MEA-01★ · MEA-03 · MEA-04 · MEA-05 · MEA-06 · MEA-07 · MEA-08
VDT (2/8)	VDT-01★ · VDT-06	VDT-02★ · VDT-03 · VDT-04 · VDT-05 · VDT-07 · VDT-08
VRF (2/16)	VRF-02 · VRF-03	VRF-01★ · VRF-04 · VRF-05 · VRF-06 · VRF-07 · VRF-08 · VRF-09 · VRF-10 · VRF-11 · VRF-12 · VRF-13 · VRF-14 · VRF-15 · VRF-16
REP (2/15)	REP-04 · REP-15★	REP-01 · REP-02 · REP-03 · REP-05 · REP-06 · REP-07 · REP-08 · REP-09 · REP-10 · REP-11 · REP-12 · REP-13 · REP-14
OPS (0/10)	—	OPS-01 · OPS-02 · OPS-03 · OPS-04 · OPS-05 · OPS-06 · OPS-07 · OPS-08 · OPS-09 · OPS-10
CNF (0/5)	—	CNF-01★ · CNF-02★ · CNF-03 · CNF-04 · CNF-05
Total	7	65

Revision notice — v1.4 → v1.5 (2026-10-20 · errata correction)

- Corresponding specification Part 1 v1.5 (errata edition — no change in assertion requirements)

- Modal forms of requirement quotations — sentences restating permissions and prohibitions of Part 1 aligned with the forms of 4.1: **PAVS-VDT-01** fail for absence of measurement ‘can be treated’ → ‘may be treated’ (composition table · verdict clause · Appendix D) · **PAVS-VDT-06** ‘cannot overwrite’ → ‘shall not overwrite’ · **PAVS-REP-04** ‘cannot raise’ → ‘shall not raise’ · **PAVS-CNF-03** ‘cannot relax’ → ‘shall not relax’ (Appendix B · Appendix D) · **PAVS-MEA-01** ‘cannot be used in scoring’ → ‘shall not be used in scoring’ (Appendix D)
- Terms — Korean term for ‘physical-hardware verification stage’ aligned with the name in Part 1 Annex A (body text and Appendix A) · Korean term for ‘self-check’ aligned (the check of 2026-08-26 · same name as in the Conformance Statement and the reference implementation) · Korean term for ‘proxy metric’ aligned (**PAVS-MEA-05**) · score ‘register’ → ‘ledger’ (**PAVS-REP-09**) · Korean term for ‘scoring system’ aligned · Part 4 ‘Implementation Profile’ → ‘Solution Profile’
- Double subject in the Appendix B NOTE resolved · estimate added to measurement provenance in Appendix D (the same four kinds as Part 1 3.5)
- Korean edition: general words replaced following the glossary of the Korean expression norm (same meaning · English wording unchanged) · the comparison target of the reproduction package named with a Korean word distinct from ‘threshold’

Revision notice — v1.3 → v1.4 (2026-10-20 · same-day textual correction)

- Clause 0 composition table — edition indications of each part deleted (an old indication remained in v1.3, giving its own edition as v1.2 and the reference implementation as v0.5 — found during comparison with the English translation); current editions = document list at the publication location
- 9.3 — edition indication of the reference implementation (v0.5) deleted · 9.4 — edition indication of the Conformance Statement deleted (current edition = publication list)

Revision notice — v1.2 → v1.3 (2026-10-20)

- 9.4 — target edition, Class C1 and the identifiers of the 7 assertions met stated; ★core status corrected to the result of the CNF-01 and CNF-02 retest (2026-10-20, not met) (inconclusive → not met); no change in the met count
- Appendix E newly established — checklist of the 72 assertions. Reason: a document that states the met count also carries the four elements of the claim form of specification 11.2 (version, class, list of assertions met and list of assertions not met) — up to v1.2 this White Paper carried only the count and references, and was judged to have omissions in the CNF-01 retest

Revision notice — v1.1 → v1.2 (2026-10-19)

This is a correction of an explanatory document; the requirements of the specification are not changed by this document (the authoritative source of the requirements is Part 1).

Change	Content	Reason
Corresponding specification	Part 1 v1.4 (editorial revision · assertion requirements identical to v1.3) · Casebook v1.2 (39 cases) · Part 3 v0.5	Alignment of bundle editions

Terms	‘edition number’ → ‘version’ · glosses at the first occurrence of item · test assertion	Defined terms kept · everyday words replaced (terminology decision of 2026-10-19)
Notation	Double angle quotation marks → single quotation marks · Korean particles attached after Latin letters, digits and closing parentheses	Korean notation norms for external documents · meaning unchanged

Revision notice — v1.0 → v1.1 (2026-10-02)

This is a correction and supplementation of an explanatory document; the requirements of the specification are not changed by this document (the authoritative source of the requirements is Part 1).

Change	Content	Reason
Reflection of specification revisions	Corresponding specification updated to Part 1 v1.3 · ITM-10 (sustainability of rated capabilities) · VRF-05 strengthened (recording of grounds for differences, prohibition of using results of mismatched models) · VDT-01 (fail for absence of measurement, retention of the not-measured count) · REP-04 (limits of self-declaration) · REP-15 (★core, indication of the physical-hardware verification stage) · description of VRF-16	The first public release was describing the wording of a previous edition
Correction of overstatements	“check execution rather than document audit” → test method types (injection 36 · audit 33 · demonstration 3) and the test-record basis · “0 gaps in clause–incident correspondence” → 3 kinds of grounds of origin · reference implementation “all assertions executable” → distinction between goal and current status	Statements stronger than the actual state
Correction of citations of institutional schemes	1 citation of a Korean scheme based on press reports deleted · scope of the EU Machinery Regulation made precise · note added that no clause recognizing simulation was confirmed for the Outdoor Mobile Robot Operational Safety Certification	Only items confirmed from primary sources kept
Correction of status	Number of cases (as of Casebook v1.1) · Part 3 v0.4 published together · Part 4 unpublished · Demonstration A	Resolution of conflicts among public documents

	reproduction package provided on request · execution time on development equipment deleted · clause ranges of the conformance classes	
Measurement dates stated	Maximum-speed case (2026-08) · retrospective audit 89% (a census of all 401 passes · 2026-05) · measured model-specification comparison (2026-08)	Prevention of misreading figures without dates
Supplementation	Conflict-of-interest notice · lineage of adjacent normative documents · origin of VRF-16 · example of a model-specification defect · 9.4 self-test · references · contact · 7 terms	Omission of information that readers need to check
Readability	Jargon and metaphors cleaned up (“anchoring” → “linking to a grounding standard”, etc.) · internal procedural wording deleted	Explanatory document for general readers

Inquiries

Inquiries, objections and requests for corrections concerning this specification and this White Paper are received at contact@immersivecast.com. The details of the objection-handling procedure (intake records, and how the outcome and its reasons are disclosed) will be set out in the next revision notice.

PAVS Part 0 White Paper · includes Appendices A–E · external citations based on the primary-source check of 2026-08-25 (partially rechecked 08-27) · previous issue 2026-10-20 (v1.4) · draft v1.5 · 2026-10-20