

Real-time governance across every AI and human media pipeline. **14 governed corpora · 1,435,331 assets · 24 scoring modules · 134 features · 17 operational surfaces.**

1 · GOVERNED DATA ESTATE

Fourteen collections across four modalities and three provenance classes. Every corpus carries its declared rating scale, origin class, asset count, integrity coefficient and verification status.

Corpus	Rating scale	Origin	Assets
wmt_da	DA (human 0–100, z-standardized)	Human	1,285,190
lsvq	DMOS 0–100	Human	38,935
videofeedback	Likert 1–4 (mean of 5 dims)	AI	33,581
nisqa_vcc	MOS (speech)	Human	20,160
fvq_20k	MOS 0–100	Human	20,000
spaq	MOS	Hybrid	11,125
koniq	MOS	Hybrid	10,373
t2vqa_db	minmax (empirical, z-derived)	AI	10,000
konvid_150k_b	Likert 1–5	Human	1,576
youtube_ugc	Likert 1–5	Human	1,380
crave_db	MOS	AI	1,227
konvid_1k	Likert 1–5	Human	1,200
live_vqc	DMOS 0–100	Human	584

Filterable by status (verified / in review / flagged) and origin; sortable by name, integrity or asset count. **Scale heterogeneity is the point:** seven distinct rating conventions are reconciled under an explicit MosScale contract, which flags non-standard scales rather than silently normalizing them.

2 · SCORING REGISTRY — 24 MODULES, 6 CATEGORIES, 134 FEATURES

Each module declares modality, maturity, layer, activation (always / conditional) and feature count. Participation is toggleable per run.

Quality — 16 modules

Module	What it measures	Feat.
prompt_alignment_fine	CLIP text–video alignment (T2V / AIGC)	24
motion_coherence	optical-flow smoothness, jerk, periodicity	10
temporal_smoothness_via_interp_artifacts	interpolation-artifact detection (≥16 fps)	8
subject_consistency_long	long-form subject identity stability	8
world_model_surprise	object persistence, world-model prediction error	5
causal_reasoning	physics-informed trajectory causal consistency	4
aesthetic_rule_of_thirds	geometric composition	4
creative_composition	shot rhythm, cut density	4
trustworthiness	noise-robustness quality signal	4
long_term_narrative	long-term coherence (DINOv2, histogram fallback)	4
semantic_consistency	CLIP frame quality and temporal consistency	3
background_consistency	SSIM background stability across shots	3
temporal_signature	flicker, judder, dropout detection	3
physics_plausibility	ghost artifacts, trajectory physics	3
aesthetic_cinematic	colour harmony, cinematic quality	3
identity_consistency	face identity across shots (ArcFace + YuNet)	1

Module	What it measures	Feat.
audio · image · text · mllm_judge	conditional, modality-gated	—

Baseline — 2 published reference scorers

- **dover_score** — DOVER aesthetic + technical UGC scorer (ICCV 2023)
- **fastvqa_score** — FAST-VQA-M efficient end-to-end VQA (ECCV 2022)

Safety, origin detection, explainability, benchmarking — 6 modules

- **multi_modal_safety** — CLIP zero-shot NSFW / violence / distress / weapon / AI-artifact detection
- **origin_detection** — RandomForest AIGC classifier over 16 signal features; a second DINOv2-based classifier cross-checks it
- **explainability** — Shapley-like attribution values per module and per metric
- **benchmark_forensic** and **benchmark_fusion** — improvement-over-raw signal, and a weighted composite from multiple gate signals

Six integrated third-party models, each licence-tracked: DOVER (S-Lab 1.0, non-commercial) · FAST-VQA (MIT) · DINOv2 (Apache-2.0) · ArcFace (MIT) · YuNet (BSD) · VideoScore (research use).

3 • EXECUTION — PLANNING, ROUTING, REPRODUCIBILITY

- **Execution Planner** (read-only, zero GPU cost) — inspect decisions before running anything: hardware routing, worker count, frame strategy, time and RAM estimates, per-module dispositions. Four profiles: fast · standard · academic · publication.
- **Runtime routing** — CPU (local) · ZeroGPU (remote runner) · T4 · L4, sequential (deterministic, reviewer-safe) or parallel. Governance rule enforced in the UI: CPU validation runs capped at $N \leq 4$.
- **Wheel lifecycle** — version base, built and promoted wheels, and automatic discovery of child Spaces with their deployed wheel version; three children currently tracked, each pinned to a build timestamp.
- **Observability** — per-run SRCC, feature count and substantiality flag across 35 tracked runs; surface health checks with latency (capabilities, modules, datasets, runs, judges); anomaly triage queue ranked by severity and confidence.
- **Per-video audit** — every scored clip in a run with its human MOS and every module/feature score, plus a predicted-versus-MOS scatter so a weak correlation is visible as a cloud rather than asserted as a number.

4 • STATISTICAL METHODOLOGY

- **Primary metric** — Spearman rank correlation (SRCC) with human MOS: rank-based, scale-invariant.
- **Also reported** — PLCC after VQEG logistic fit · KRCC · RMSE.
- **Uncertainty** — 95% bootstrap confidence intervals, 1,000 resamples.
- **Multiplicity** — Benjamini-Hochberg FDR correction, q-values across features.
- **Calibration** — reliability over 10 prediction-quantile bins; mean absolute calibration error reported *including when it is poor*.
- **Integrity** — MOS-file SHA-256 with a verification endpoint; video-set fingerprint over a content manifest, so the exact asset set is checkable without byte-hashing every file.
- **Replayability** — git commit + git-diff size + seed + config hash + calibration hash + requirements hash + MOS hash + execution plan. A diff size of 0 certifies the run reflects committed code.

5 • CROSS-DATASET GENERALIZATION — AN INTEGRITY INSTRUMENT

Per-feature |SRCC| heatmap across **134 features × 12 datasets**, with mean, standard deviation and an inversion flag per feature.

Headline finding: 99 sign inversions — features whose correlation with human judgment flips sign between corpora. The platform names this what it is: *a scientific integrity flag*, surfaced rather than hidden. The best single feature reaches |SRCC| 0.512 (gate.sharpness); most sit between 0.05 and 0.33, with σ up to 0.195.

Implication: a predictor validated on one dataset may be anti-correlated on another. This is measurable, and it is measured.

6 • LLM-AS-JUDGE — AUDITING THE AUDITORS

Multi-provider judging (OpenAI, Anthropic) with an inter-judge leaderboard reporting SRCC, PLCC and MAE against human MOS.

Judge	Corpus	n	SRCC	PLCC	MAE
gpt-4o-mini	t2vqa_db	20	0.188	0.007	49.56
gpt-4o-mini	youtube_ugc	50	0.042	0.013	1.02
claude-haiku-4.5	fvq_20k	20	-0.064	-0.084	46.71
DOVER (conventional metric)	lsvq	—	~0.86	—	—

A reproducible metric outperforms a fragile LLM judge by more than an order of magnitude, and one judge correlates *negatively* with human ratings. The platform states this plainly rather than omitting it.

7 • REVIEWER-SAFE REPORT GENERATION

Three report kinds — **reproducibility appendix · draft paper · architecture paper** — each exportable as Markdown, LaTeX, DOCX and PDF, assembled from real run artifacts.

Governing rule, enforced in the generator: **every value traces to a manifest, metrics file or audit; missing data is stated, never invented**.

A generated appendix carries seven sections: reproducibility metadata · dataset and integrity · statistical methodology · results with 95% CIs and BH q-values · calibration analysis · cross-dataset robustness · **mandatory limitations and caveats**.

The limitations section is not decorative. A live example declares: staging modules producing no features are excluded from stable claims; a non-standard MOS scale is flagged by contract; GPU-quota-blocked large runs are marked in-review rather than reported as final; and rank-based SRCC claims are distinguished from absolute-scale PLCC/RMSE claims whose validity depends on the reported calibration error.

8 • ATTRIBUTION AND LICENSING

Every dataset and model cited to its original authors, venue and licence — **12 datasets, 6 models** — spanning QoMEX, CVPR, ECCV, ICCV, EMNLP, ACM MM, IEEE TIP, IEEE Access, Odyssey and WMT. Licences tracked individually: CC BY, CC-BY-NC-SA-4.0, Apache-2.0, MIT, BSD, S-Lab 1.0, and per-database research-use terms.

Stated principle: scientific credit is a reviewer-safe requirement.

9 • INTERFACE AND ACCESSIBILITY

- **17 operational surfaces** — Overview · Datasets · Modules · Anomalies · Runtime · Plan · Runner · Wheel · Benchmark · Leaderboard · Cross-dataset · Reports · LLM Judge · Audit · References · Design System · Multi-Video.
- **Design system** — 18 themes across scientific, security/compliance, finance/data, cinema and editorial registers, exposed through a /ui-schema endpoint.
- **Internationalization** — five languages: English, Spanish, French, Arabic (right-to-left) and Chinese.
- **Multi-Video workbench** — batch upload streaming to a shared bucket, GPU-runner scoring at parallelism 248, and feature-by-feature comparison via radar, heatmap, correlation and ranking views across 16 module groups.

10 • ENGINEERING DISCIPLINE

Over 90% test coverage executing in under 30 seconds · computation-level audit logging of every intermediate step · semantic versioning with full backward compatibility · automated CI/CD that interrupts the build on test failure and issues explicit go/no-go release decisions · PyPI packaging · Child-Space version discovery and promotion · Zenodo DOI archival for permanent citability (example: doi.org/10.5281/zenodo.22698801)

THREE PUBLISHABLE RESULTS ALREADY IN THE DATA

- 1 • **LLM judges are unreliable evaluators of perceptual quality.** SRCC 0.188 / 0.042 / -0.064 against human MOS, versus roughly 0.86 for a conventional metric on a comparable task.
- 2 • **Ninety-nine sign inversions across 134 features and 12 datasets.** A predictor's correlation with human judgment flips sign between corpora — generalization failure made measurable rather than argued.
- 3 • **Small-sample correlation inflation, demonstrated on the platform's own runs.** SRCC collapses from 0.670 to 0.316 (videofeedback), 0.638 to 0.209 (konvid_1k) and 0.562 to 0.151 (lsq) between N=5 and N=50. A figure reported at N=5 would be roughly four times too high — the platform proves its own thesis on its own data.