NEURIPS 2026 · PARIS · EVALUATIONS & DATASETS TRACK

RASS

Risk-Audited Budget Selection for Compact NeRF Benchmark Subsets

Ahmad AlMughrabi[1] Flavià Serret[1] Ricardo Marques[2†] Petia Radeva[1,3†]

[1]Universitat de Barcelona [2]Universitat Pompeu Fabra [3]Institut de Neurosciències, Barcelona [†]Equal supervision

Code & data
github.com/GCVCG/RASS

Abstract. NeRF evaluation trains one model per scene. RASS selects the smallest subset size whose random draws meet declared PSNR, SSIM and LPIPS mean and distribution tolerances at a rate certified by a **Wilson lower bound**. On 3,521 Nutrition5k scenes it yields **RASS-48** for screening and **RASS-96** for reporting.

01 INTRODUCTION Benchmark cost and risk

3,521 NeRF models trained per method

512 H100 GPUs (maximum)

≈73× Training CO_2 reduction per method (RASS-48)

Agreement on a small subset can be spurious:

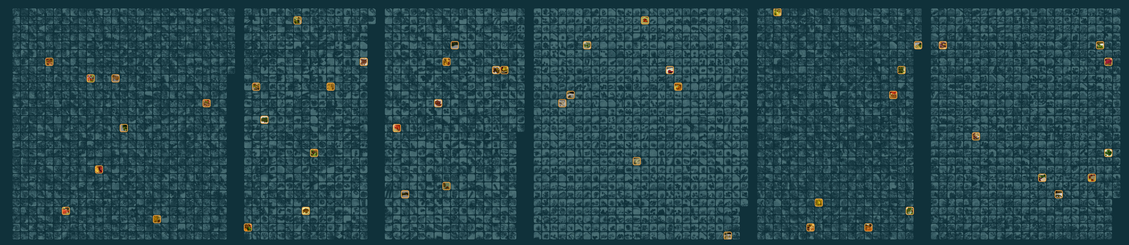
- **distributional mismatch**: equal means, different tails;
- **cross-metric mismatch**: one metric agrees, others deviate;
- **sampling variability**: a passing draw may not replicate.

Question. What is the smallest subset preserving chosen benchmark summaries under explicit fidelity constraints?

48 scenes instead of 3,521

certified by a statistical audit: ≈73× fewer NeRF models per method

Each tile is one audited scene, grouped by regime; in color: RASS-48, 8 scenes per regime (right).



02 METHOD Risk-audited budget selection

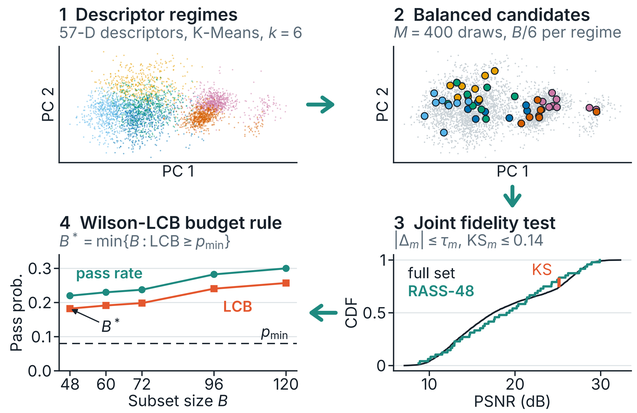


Figure 1. RASS pipeline on Nutrition5k, panels in clockwise order (panel 2: RASS-48). The audit uses only **existing evaluation logs**.

03 BASELINES Candidate generators

Table 1. Pass rates over 400 draws of 48 scenes; ↑: higher is better.

Generator	All means ↑	All KS ↑	All six ↑	Wilson LCB ↑
Uniform random	0.148	0.628	0.138	0.107
Balanced (RASS)	**0.238**	**0.730**	**0.220**	**0.182**

Smallest certified size: facility location 36 (reimplementation unstable), balanced 48 (36 on the extended grid), random 48; k-center none ≤ 96.

04 RESULTS Audit at 48 scenes

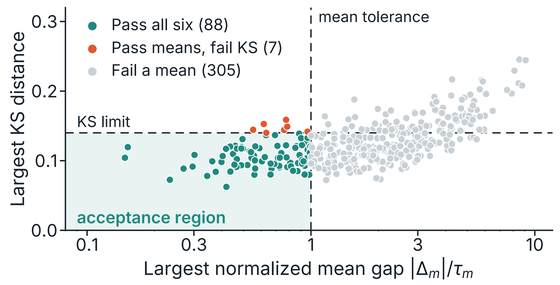


Figure 2. Each point is one of 400 balanced draws of 48 scenes; draws inside the shaded box pass all six constraints: 88 (rate 0.220, **Wilson LCB 0.182 ≥ p_{min} = 0.08**).

05 RESULTS Cost–reliability frontier

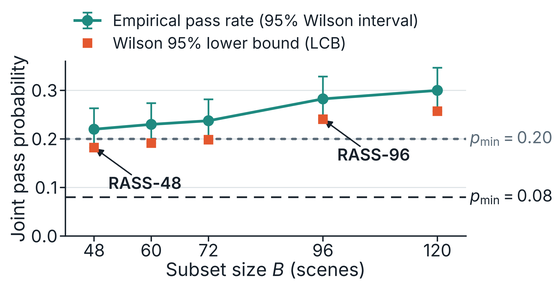


Figure 3. Pass rate with 95% Wilson intervals (400 draws per budget). RASS selects the smallest evaluated budget whose **lower bound (red)** reaches p_{min}: **48 for 0.08, 96 for 0.20**.

06 ROBUSTNESS Sensitivity analysis

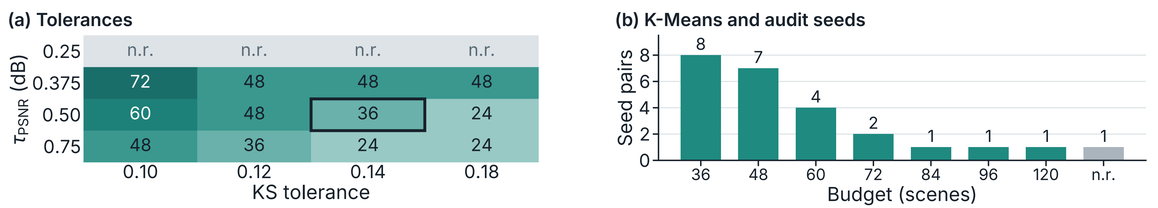


Figure 4. Recommended budget at p_{min} = 0.08 on the extended grid. (a) **Stricter tolerances require larger budgets** ($\tau_{SSIM} = \tau_{LPIPS} = 0.02\ \tau_{PSNR}$; n.r.: not reached by 120; outlined: defaults, 36 on this grid, 48 on the reported grid). (b) Over 25 seed pairs, the budget ranges from **36 to 120 scenes** (median 48).

07 VALIDATION Global fidelity of RASS-48

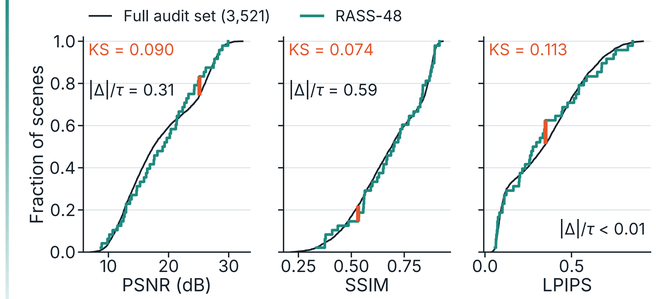


Figure 5. Per-scene Zip-NeRF scores; red: KS distance. All mean gaps ($|\Delta|/\tau < 1$) and KS distances (< 0.14) are **within tolerance**.

08 GENERALIZATION Methods and datasets

Table 2. Constraints passed.

Methods	RASS-48 ↑	RASS-96 ↑
3	14/16	**16/16**
4	19/22	**22/22**
5	23/28	26/28

Table 3. Ranking preserved.

Metric	All 400 ↑	88 passing ↑
PSNR	95.1%	95.5%
SSIM	84.2%	94.9%
Both	80.1%	91.0%

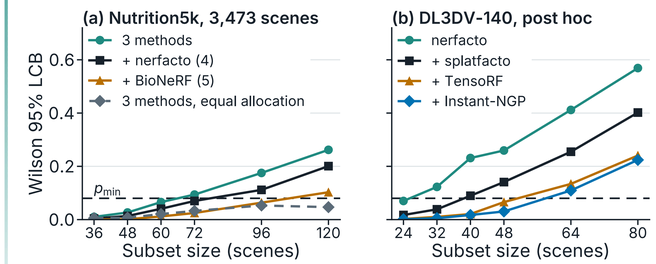


Figure 6. Wilson LCB of joint audits, with proportional allocation unless marked. **Larger method sets require larger budgets**: LCB ≥ 0.08 at 72, 96, 120 scenes for 3–5 methods (Nutrition5k); 32, 40, 64, 64 for 1–4 (DL3DV, **frontier post hoc**). Tables: 3,473 scenes; ranking is diagnostic.

09 CONCLUSIONS Released subsets and usage

RASS-48
Screening, p_{min} = 0.08

RASS-96
Reporting, p_{min} = 0.20

Protocol: **train and evaluate on the subset**, compare with the released logs, and report the subset version.

Fidelity is **certified globally**; per-regime and leaderboard claims require the full benchmark. Project page: amughrabi.github.io/RASS

References. Nutrition5k: Thames et al., CVPR 2021 · Pre-NeRF 360: AlMughrabi et al., IbPRIA 2025 · Zip-NeRF: Barron et al., ICCV 2023 · DL3DV-10K: Ling et al., CVPR 2024 · Wilson score interval: Wilson, JASA 1927. Scene images: Nutrition5k.

Funding. TRIVISION (PID2025-173459NB-C21), PULSAR (PID2025-173459NB-C22) and PID2022-141566NB-I00, financed by MICIU/AEI/10.13039/501100011033 and by FEDER/UE; AIA2025-163919-C51 (AEI, MICIU); CNS2022-135480 (MICIU/AEI, FEDER, NextGenerationEU/PRTR); ICREA Academia'2022 (Generalitat de Catalunya); the FPI doctoral fellowship; and RES resources at BSC MareNostrum5 (IM-2025-3-0008).

Universitat de Barcelona · upf. Universitat Pompeu Fabra Barcelona · BSC Barcelona Supercomputing Center, Centro Nacional de Supercomputación