

# RASS

**48 scenes instead of 3,521,  
backed by a statistical audit**

Universitat de Barcelona · Universitat Pompeu Fabra ·  
Institut de Neurosciències, Barcelona



# Full NeRF benchmarks are too costly to rerun

## 3,521

NeRF trainings for every ablation on the full benchmark

## 512

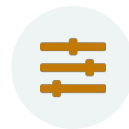
H100 GPUs, at most, to produce the full logs once

## Ad hoc subsets can look right for the wrong reasons



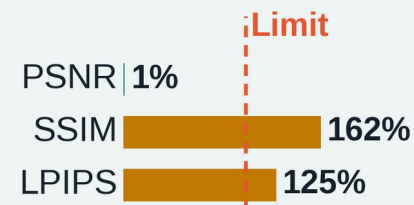
### Tail shift

A real 48-scene draw: every mean matches, but its LPIPS distribution (amber) drifts



### Metric leakage

Another real draw: PSNR matches, but SSIM and LPIPS overshoot their allowed gaps



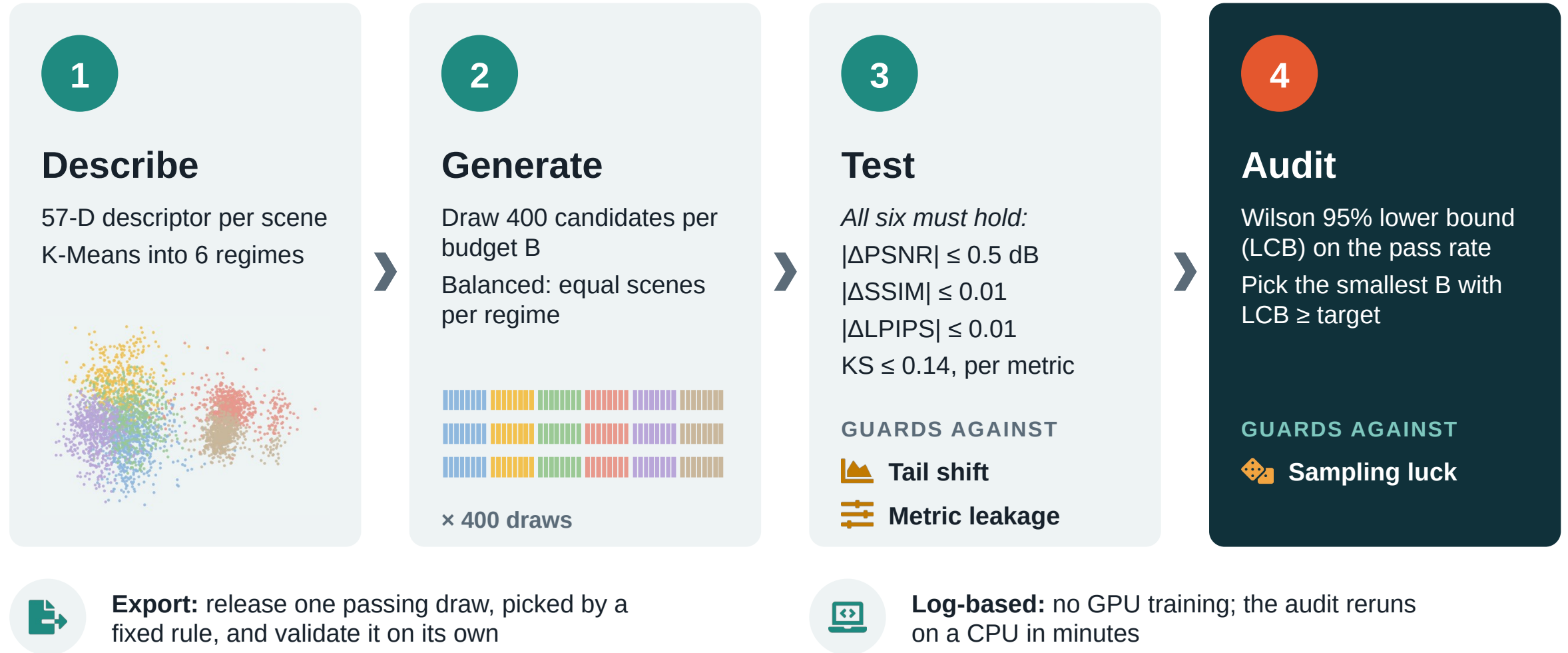
### Sampling luck

One subset passes by chance; the next draw does not

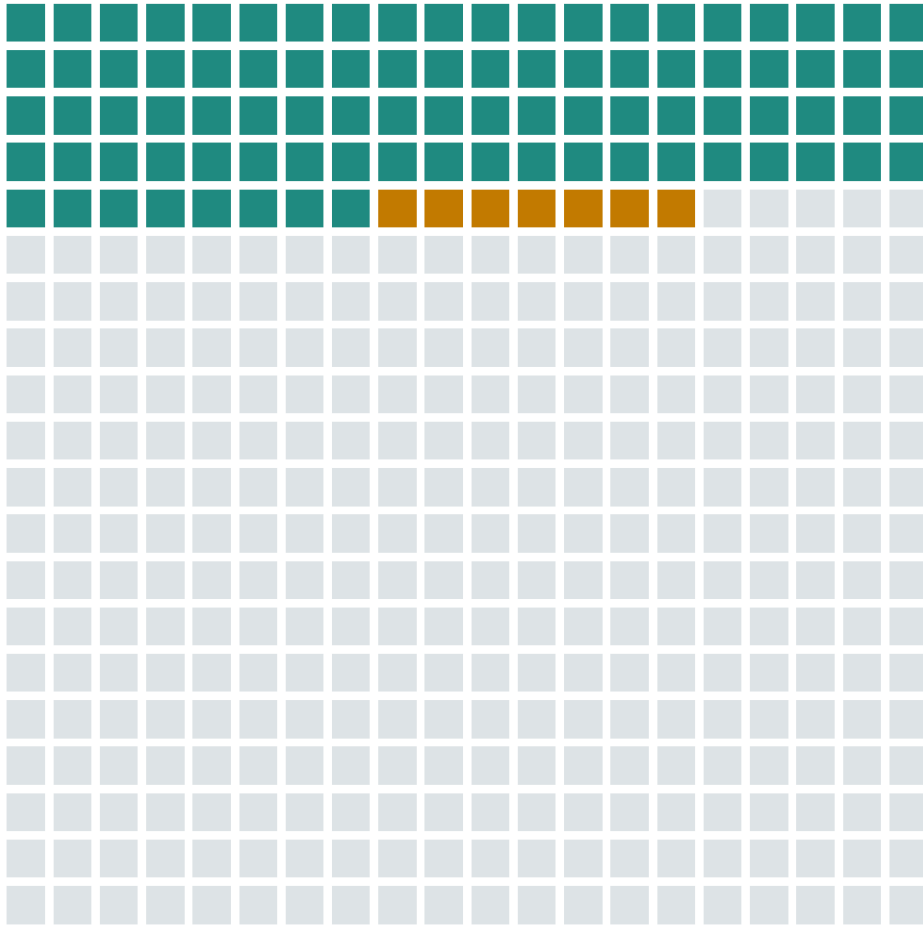


**Our question:** what is the smallest subset that preserves the full-benchmark summaries, under explicit fidelity constraints?

# RASS audits the procedure, not one lucky subset



# At 48 scenes, even the pessimistic rate clears the target



- Pass all six constraints (88)
- Means match, KS fails: tail shift caught (7)
- Fail a mean constraint (305)

Pass rate  $88 / 400 = 0.22$

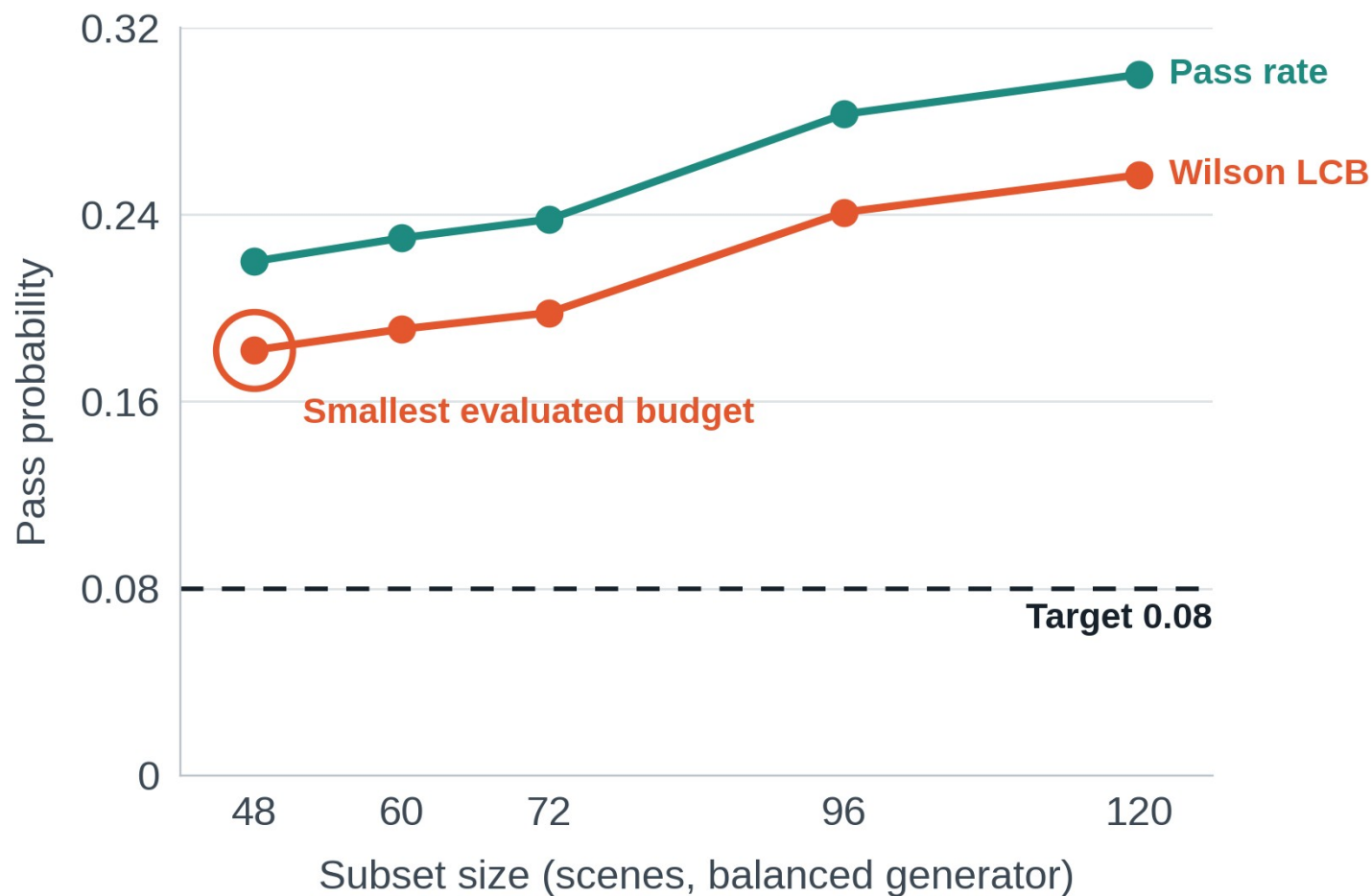
Wilson 95% LCB  $0.182 \geq$  target 0.08

✓ 48 scenes certified

**The target is a yield:** on average, at most 12.5 cheap log checks find a passing subset.

The 400 real candidate draws at 48 scenes

# The frontier trades scenes for reliability



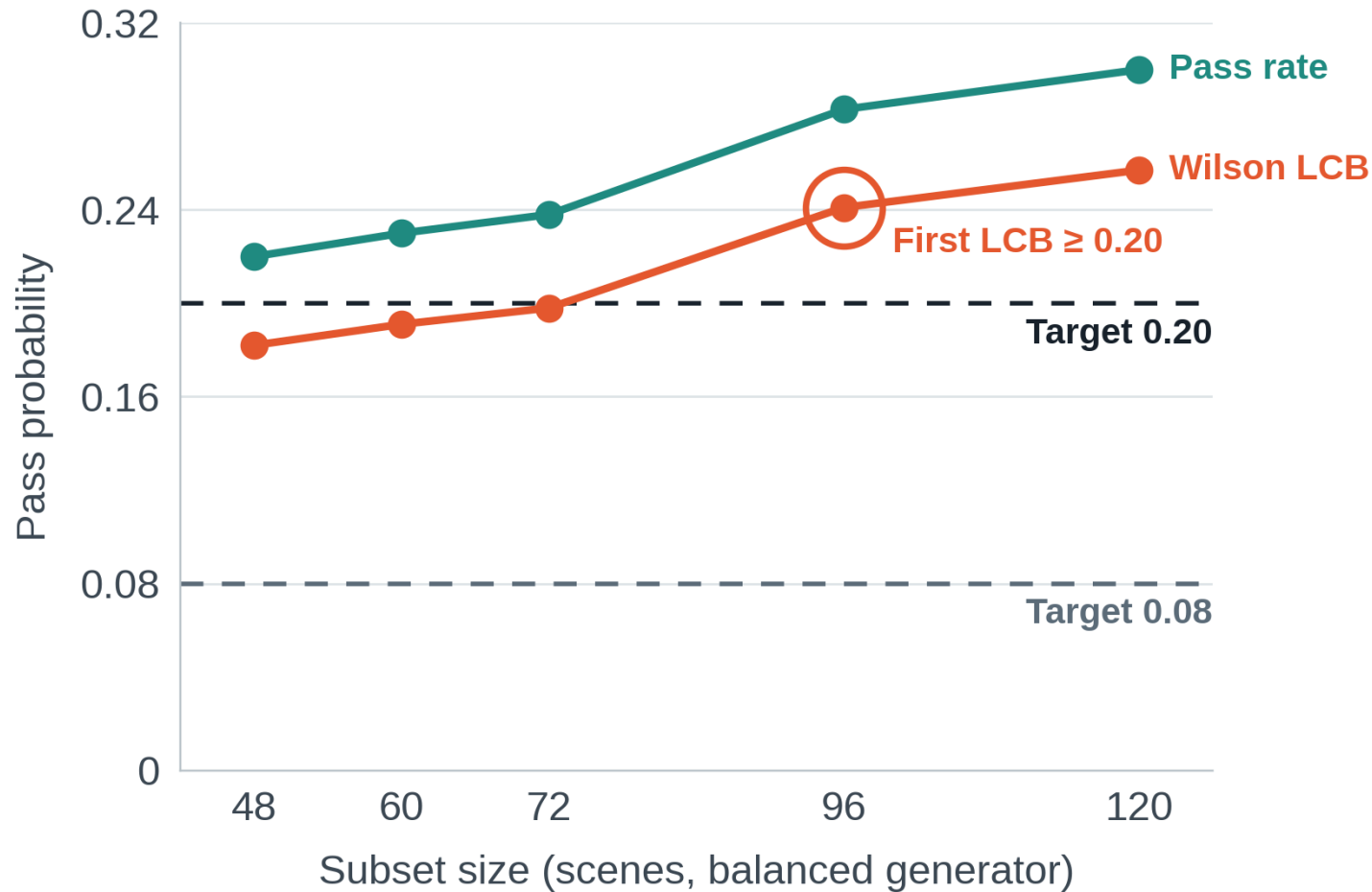
Target 0.08

## RASS-48

48 scenes · LCB 0.182

*Screening & ablations*

# The frontier trades scenes for reliability



Target 0.08

## RASS-48

48 scenes · LCB 0.182

*Screening & ablations*

Target 0.20

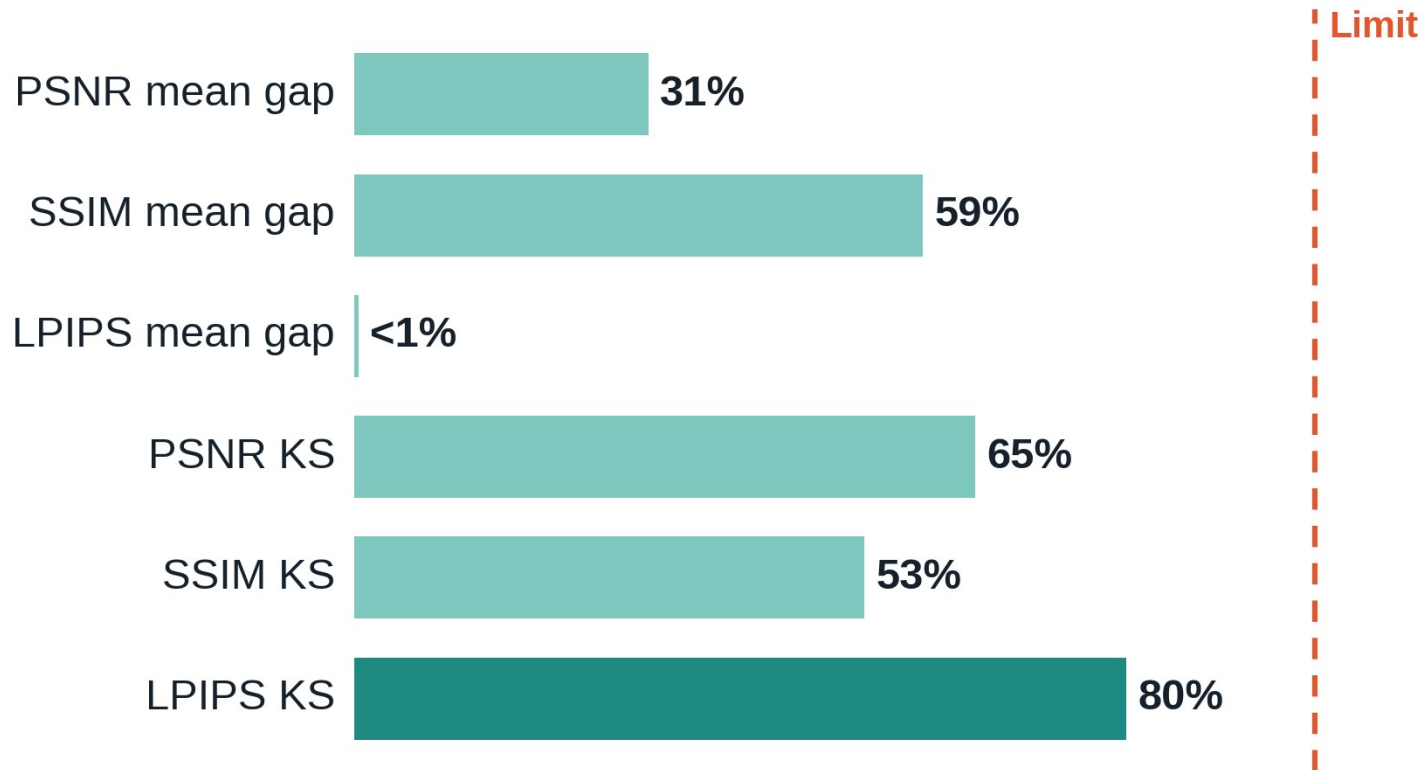
## RASS-96

96 scenes · LCB 0.241

*Compact reporting*

# RASS-48 passes the global audit with 73× fewer scenes

Share of each tolerance that RASS-48 uses



# 73×

**fewer scenes: 3,521 to 48**

8 scenes from each of the  
6 descriptor regimes



**Passes all six checks**

# One audit covers more methods, generators, and datasets



**RASS-96 passes the 3- and 4-method audits**

Constraints passed in joint audits

3 methods + nerfacto

**RASS-48**    **14/16**    **19/22**

**RASS-96**    **16/16**    **22/22**

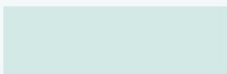
So RASS-96 is our subset for reported results.

3 methods: Zip-NeRF, Feature-Splatting, Instant-NGP (no LPIPS)



**Any candidate generator plugs in**

Smallest certified size (scenes)

Facility loc. (unstable)  36

Balanced  48

Random  48

**✗ k-center fails up to 96**

Balanced also passes at 36 on an extended grid.



**The protocol transfers to DL3DV**

**64 / 140**

DL3DV scenes certified for four methods (32 for one)

Contract re-declared for DL3DV; proportional allocation, added post hoc.

# One audit covers more methods, generators, and datasets



RASS-96 passes the  
3- and 4-method audits

Constraints passed in joint audits

3 methods + nerfacto

RASS-48 14/16 19/22

RASS-96 16/16 22/22

So RASS-96 is our subset  
for reported results.

3 methods: Zip-NeRF, Feature-Splatting,  
Instant-NGP (no LPIPS)



Any candidate  
generator plugs in

Smallest certified size (scenes)

Facility loc.  
(unstable) 36

Balanced 48

Random 48

✗ **k-center fails up to 96**

Balanced also passes at 36 on an extended  
grid.



The protocol  
transfers to DL3DV

64 / 140

DL3DV scenes certified for four  
methods (32 for one)

Contract re-declared for DL3DV;  
proportional allocation, added post hoc.

# One audit covers more methods, generators, and datasets



RASS-96 passes the  
3- and 4-method audits

Constraints passed in joint audits

3 methods + nerfacto

RASS-48 14/16 19/22

RASS-96 16/16 22/22

So RASS-96 is our subset  
for reported results.

3 methods: Zip-NeRF, Feature-Splatting,  
Instant-NGP (no LPIPS)



Any candidate  
generator plugs in

Smallest certified size (scenes)

Facility loc.  
(unstable) 36

Balanced 48

Random 48

✗ k-center fails up to 96

Balanced also passes at 36 on an extended  
grid.



The protocol  
transfers to DL3DV

64 / 140

DL3DV scenes certified for four  
methods (32 for one)

Contract re-declared for DL3DV;  
proportional allocation, added post hoc.

# Certify the procedure, not one lucky subset

## RASS-48

### *Screening & ablations*

Smallest evaluated budget with  $LCB \geq 0.08$

## RASS-96

### *Compact reporting*

$LCB \geq 0.20$ ; passes the 3- and 4-method joint audits

**How to use:** train only on RASS-48 or RASS-96, compare with our released logs on the same scenes, and report the subset version.



[github.com/GCVCG/RASS](https://github.com/GCVCG/RASS)



[kaggle.com/datasets/ahmadalmughrabi/rass-nerf-benchmark-artifact](https://kaggle.com/datasets/ahmadalmughrabi/rass-nerf-benchmark-artifact)



Code & data

## Thank you