NEURIPS 2026 / EVALUATIONS AND DATASETS

# GenScale

## A Benchmark for Relative Object Scale in Image Generation and Editing

**Lingxiao Li[1], Max Whitton[1], Ledell Wu[2], and Boqing Gong[1]**

[1] Boston University [2] Creatify AI

**How well do generative models preserve real-world object sizes?**

# Photorealistic images can get object sizes wrong



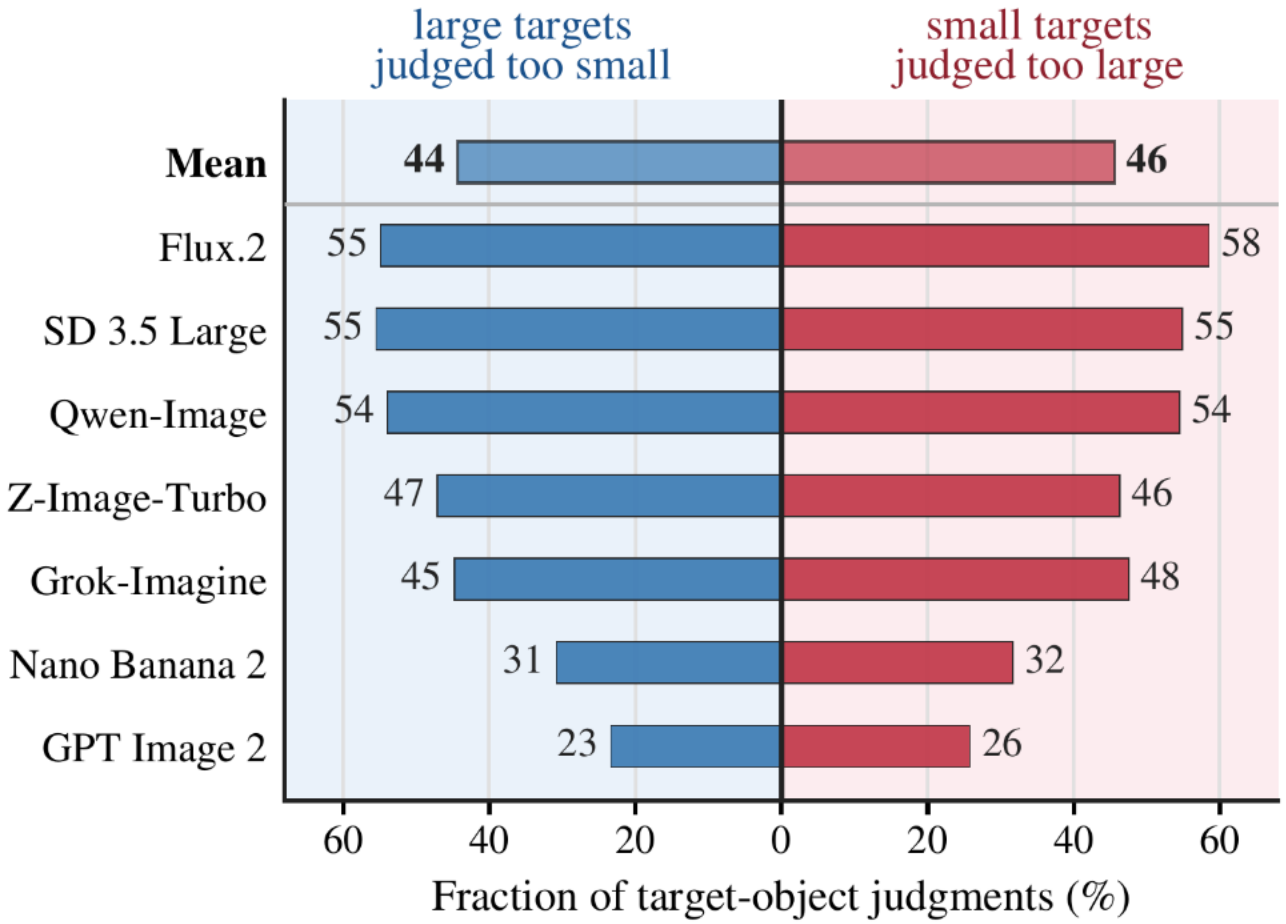
Qwen-Image 2512



SD3.5-Large

**Both generators make the banana implausibly large relative to the bicycle.**

# Generators compress relative object sizes




**Mean regression**

## Large objects

Often rendered too small

## Small objects

Often rendered too large

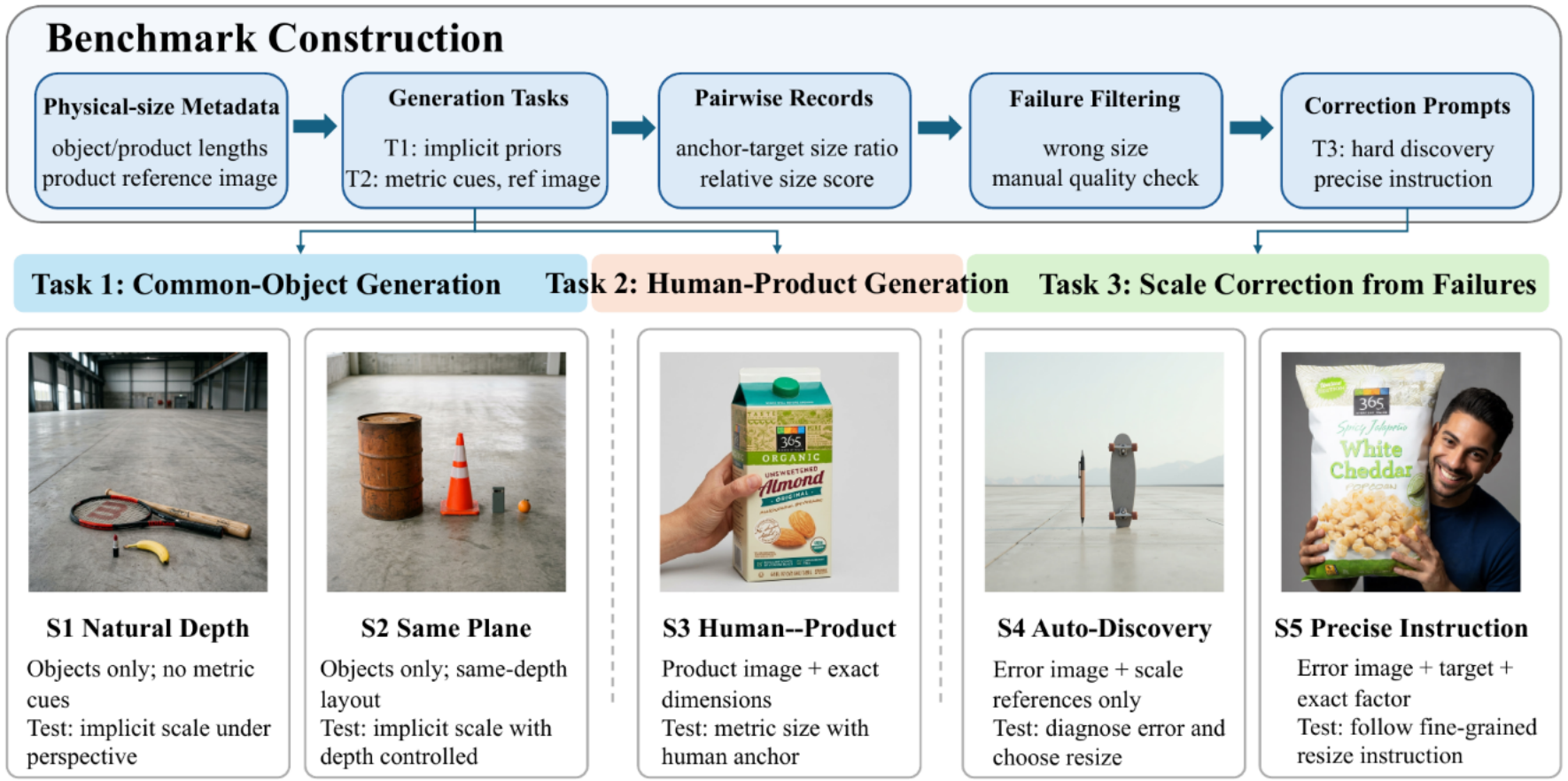
**The pattern appears across all seven generators tested.**

# Prior benchmarks cover related capabilities

| | Broad capabilities | | | Scale-focused evaluation | | |
|---|---|---|---|---|---|---|
| **Benchmark** | **Composition** | **Physics** | **Spatial** | **Object–object** | **Human–product** | **Scale correction** |
| GenEval | ✓ | × | ✓ | × | × | × |
| T2I-CompBench++ | ✓ | × | ✓ | × | × | × |
| GenAI-Bench | ✓ | × | × | × | × | × |
| PhyBench | × | ✓ | × | × | × | × |
| GenSpace | × | × | ✓ | ✓ | × | × |
| **GenScale** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

**GenScale evaluates real-world size ratios in generation and correction.**

# GenScale tests three relative-scale tasks




**900** image-level entries **1,643** anchor-target pairs **5** scenarios

# The scale judge agrees closely with human raters




**Five-point scale rubric**

| | |
|---|---|
| **1–2** | Target too small |
| **3** | Plausible scale |
| **4–5** | Target too large |

**527 pairs from 285 images, scored by nine raters**

**0.832**

Pearson correlation with human consensus

**96.77%**

Agreement within one score level

# Task 1 · Common-object relative scale

| Model | S1 · Natural depth | | | S2 · Same plane | | | S1 + S2 · All scenes | | |
|---|---|---|---|---|---|---|---|---|---|
| | Err. ↓ | Plaus. ↑ | MR ↓ | Err. ↓ | Plaus. ↑ | MR ↓ | Err. ↓ | Plaus. ↑ | MR ↓ |
| Nano Banana 2 | 0.51 | 65.1 | 15.8 | 0.69 | 47.8 | 47.9 | **0.60** | **56.5** | 31.8 |
| GPT-Image-2 | 0.78 | 51.1 | 12.1 | 0.63 | 52.8 | 35.8 | 0.71 | 51.9 | **24.0** |
| Z-Image-Turbo | 0.64 | 53.4 | 35.7 | 0.88 | 38.1 | 55.7 | 0.75 | 46.2 | 45.1 |
| Grok Imagine | 0.60 | 58.7 | 26.9 | 0.98 | 34.7 | 61.1 | 0.79 | 46.6 | 44.1 |
| Qwen-Image 2512 | 0.70 | 53.1 | 37.5 | 1.10 | 27.0 | 67.6 | 0.91 | 39.8 | 52.9 |
| FLUX.2 | 0.83 | 45.5 | 40.4 | 1.21 | 23.3 | 69.4 | 1.03 | 33.9 | 55.5 |
| SD3.5-Large | 0.96 | 36.7 | 46.7 | 1.19 | 24.5 | 70.0 | 1.07 | 31.0 | 57.7 |

Err. = mean scale error, Plaus. = plausible pairs (%), MR = mean regression errors (%).

**Mean regression is more frequent on the same plane for every model tested.**

# Task 2 · Human-anchored product scale

Models receive a product image, its physical dimensions, and a human reference.

| Model | Valid images / pairs | Scale error ↓ | Plausible (%) ↑ | Severe (%) ↓ |
|---|---|---|---|---|
| GPT-Image-2 | 294 / 294 | **0.231** | **78.9** | **2.0** |
| Nano Banana 2 | 295 / 295 | 0.268 | 75.6 | 2.4 |
| Seedream 4.5 | 295 / 295 | 0.302 | 74.9 | 5.1 |
| Qwen-Image-Edit-2511 | 297 / 297 | 0.327 | 71.0 | 3.7 |
| FLUX.1 Kontext-dev | 291 / 291 | 0.423 | 63.2 | 5.5 |
| SD3.5-Large + IP-Adapter | 269 / 269 | 0.595 | 52.8 | 12.3 |

Severe errors are ratings at either extreme of the five-point rubric.

**When models get product size wrong, they more often make it too large.**

# Precise instructions help editors correct scale

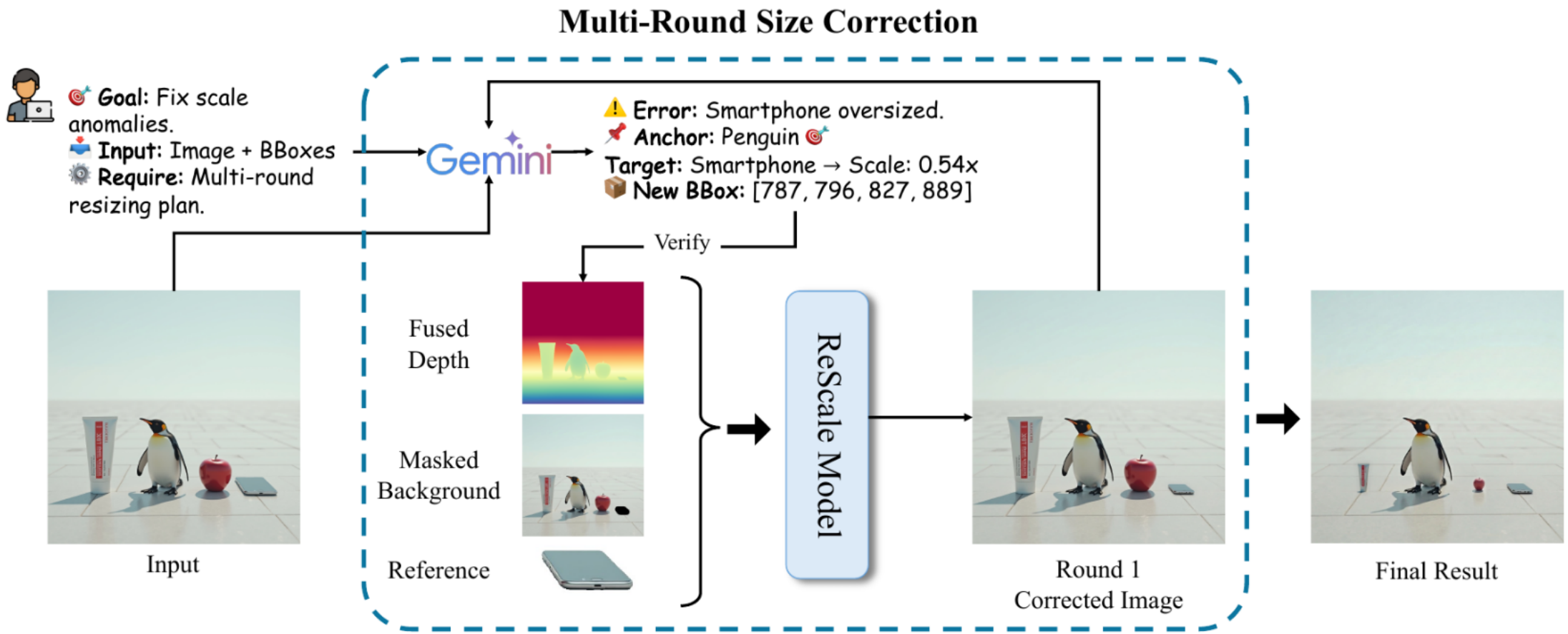
Task 3 scale error on scorable outputs. Lower is better.

| Editor | S4 Auto-discovery | S5 Precise instruction |
|---|---|---|
| GPT-Image-2 | 0.96 | 0.41 |
| Nano Banana 2 | 1.01 | 0.93 |
| FLUX.1 Kontext-dev | 1.29 | 0.96 |
| Qwen-Image-Edit-2511 | 1.24 | 1.00 |
| Seedream 4.5 | 1.17 | 1.10 |
| SD3.5-Large + IP-Adapter | 0.86 | 1.31 |

SD3.5-Large + IP-Adapter has 14 scorable outputs in S4 and 59 in S5.

**Five of six editors achieve lower scale error with precise resize instructions.**

# Rescale uses feedback to plan local edits




**Estimate the resize factor, reinsert the object, and verify the edited scale.**

# Local edits improve relative-scale plausibility



Before



After Rescale

**1.258 → 0.548**

T3 scale error on 93 unique matched source pairs

**+21.0 points**

T1 plausible rate across 6,865 matched pairs

**The frisbee anchors the edit. Rescale shrinks the oversized watch.**

# GenScale measures where relative scale breaks

**Benchmark**

Three tasks probe relative size in generated and edited images.

**Evaluation**

Physical reference lengths ground the human-calibrated scale judge.

**Correction**

Rescale turns scale feedback into a localized edit.

**Paper, benchmark, and code**

**lingxiao-li.github.io/genscale/**