# FROZEN SCENES, SHIFTING WINNERS: CONFIGURATION FRAGILITY IN TEXT-TO-3D EVALUATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Can a text-to-3D leaderboard change when every generated scene stays fixed? We audit this question for rendered-image evaluation, where camera settings and caption wording become part of the measurement protocol. Across 300 frozen scenes from six generators, we vary eight render and caption factors for 19 alignment evaluators plus one perceptual-quality control, then test four targeted scene degradations. Peak configuration variance exceeds between-generator variance for 17/19 alignment evaluators, with prompt-bootstrap lower bounds above 1 for 11/19. Rankings are more stable than scores, yet 18/19 evaluators change their point-estimate winner under some configuration. Pairwise protocol margin envelopes show which comparisons keep their direction across the tested settings. Selected pairs have opposite pointwise intervals, but no reversal survives simultaneous inference over the full search. Thus the observed winner changes are descriptive, not confirmed changes in generator superiority. Sensitivity remains separate: no evaluator, even the prompt-free control, exceeds 67% tie-adjusted directional discrimination on layout scrambling, which is diagnostic rather than human-validated ground truth. The audit separates score stability, decision uncertainty, and targeted sensitivity, and recommends reporting *(generator, score, card ID)* with protocol-dependent comparisons and selection-aware uncertainty.

## 1 INTRODUCTION

A text-to-3D scene paper can change its reported winner without changing any generated scene. A common evaluation recipe renders one view, wraps the prompt in a caption template, and asks a vision-language model whether they match. Camera, lighting, and wrapper wording then become part of the evaluator. The scientific question is which scene-generator comparisons remain credible when those choices are made explicit.

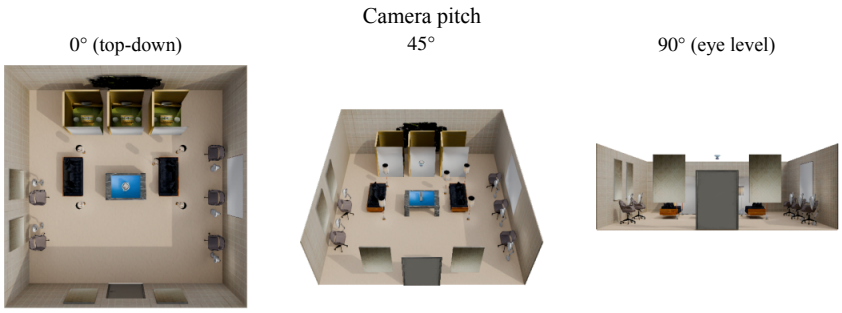



Figure 1: **The artifact stays fixed, but the measurement changes.** Three camera pitches reveal different evidence from one frozen room. Our audit holds all 300 scene artifacts fixed and varies the rendering or caption protocol. Winner changes are measured across all 50 prompts, not inferred from this pictured room, and are separated from statistical confirmation in Section 4.3.

This matters because generated indoor scenes are becoming useful, navigable artifacts for design, media, and embodied AI (Yang et al., 2024; Lin & Mu, 2024; Berdoz et al., 2026). As generators



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

improve, a metric that responds to an incidental view or wrapper can steer comparisons toward the wrong model. Building on SceneCritic's view-dependent VLM-judge findings (Sengupta et al., 2026) and 3D-DefectBench's pipeline audit for human-referenced defect detection (Zhao et al., 2026), we ask a narrower decision question: within a named evaluator, how much do plausible render and caption settings move scores and generator comparisons on the same fixed scenes?

We audit this question under a fixed-artifact counterfactual design. Holding all generated scenes fixed, we vary one measurement-protocol factor at a time across 19 alignment evaluators and 1 perceptual-quality control. Following measurement theory (Zhao et al., 2024; Ross et al., 2024), this design separates reliability under configuration choices from sensitivity to targeted scene changes. We call a change in the unique top-ranked generator a *same-scene point-estimate winner reversal* when only the evaluator configuration changes. The term describes the reported decision, not statistically established generator superiority.

The audit gives three main findings. First, peak configuration variance exceeds between-generator variance for 17 of 19 alignment evaluators, with bootstrap lower bounds above 1 for 11. Second, rankings are more stable than scores, but 18 evaluators change their point-estimate winner. Selected generator pairs even yield opposite pointwise intervals, although no reversal survives the full-search simultaneous band. Third, rank stability does not imply scene understanding: no evaluator exceeds $67\%$ tie-adjusted discrimination on layout scrambling. Pairwise protocol margin envelopes connect these observations by identifying which sampled comparisons keep their direction across the declared configuration set.

Our contributions are:

- **A fixed-artifact audit design** covering 300 frozen scenes, six generators, 19 alignment evaluators plus one control, eight render and caption factors, and four targeted degradations.
- **A decision analysis** that separates score drift, descriptive winner changes, pairwise protocol margins, pointwise uncertainty, and full-search simultaneous inference.
- **A reporting contract** centered on a versioned Evaluator Configuration Card, so each reported score can be tied to the render and caption choices behind it.

The practical outcome is an audit before a superiority claim: declare the evaluator protocol, inspect method differences over a stated configuration set, quantify paired uncertainty, and test sensitivity to scene properties independently.

## 2 Related Work

**Rendered-image evaluation for text-to-3D scenes.** Text-to-scene generators span procedural construction, diffusion layouts, VLM optimization, and reasoning pipelines (Yang et al., 2024; Tang et al., 2024; Sun et al., 2025; Berdoz et al., 2026). Their alignment evaluation often reduces a 3D scene to one rendered image and one text query: Holodeck uses top-down CLIPScore (Yang et al., 2024), Holodeck 2.0 scores perspective renders (Bian et al., 2025), and HSM adds BLIP-2, LongCLIP, and VQAScore (Pun et al., 2026). Because rendering settings and caption wrappers are often under-specified, the metric name alone does not define the measurement.

**Metric audits and measurement evidence.** CLIPScore made reference-free image-text scoring common (Hessel et al., 2021), but CLIP-like encoders struggle with relations, counting, and viewpoint changes (Yuksekgonul et al., 2023; Paiss et al., 2023; Hou et al., 2025). Newer metrics use question answering or VLM scoring (Hu et al., 2023; Lin et al., 2024), while audits reveal score artifacts, prompt sensitivity, and controlled-degradation blind spots (Schaeffer et al., 2023; Ross et al., 2024; Charles et al., 2025). AutoMetrics is closest in spirit because it separates stability under same-quality perturbations from sensitivity to worse-quality perturbations (Ryan et al., 2026). RLiable supplies uncertainty-aware comparison tools (Agarwal et al., 2021), and Evaluation Cards records evaluation-run provenance for comparability (Ghosh et al., 2026). Our card records render and caption choices for published vision evaluators on frozen 3D scene artifacts, and our audit traces their effects to generator decisions.

**Fine-grained 3D evaluation and closest boundaries.** Object and mesh benchmarks use multi-view aggregation or learned evaluators (He et al., 2023; Zhang et al., 2025a; Jia et al., 2026), while scene benchmarks improve spatial and trajectory-based evidence (Tam et al., 2026; Duan et al., 2025). SceneCritic is the closest same-artifact instability study: it reports repeated-call and view-dependent score variance and method-ranking reversals for one Gemini-2.5-Pro judge, then proposes a human-validated symbolic floor-plan evaluator (Sengupta et al., 2026). 3D-DefectBench studies the full VLM-based defect-detection pipeline, crossing judge, camera protocol, visual input, and prompt schema against human fine-grained defect labels (Zhao et al., 2026). Our contribution is different in scope. We vary render and caption settings within each of 19 alignment evaluators, compare configuration variance with between-generator signal, add targeted scene degradations, and report pairwise protocol envelopes with selection-aware uncertainty. Thus the paper supports a reporting contract for score comparability rather than a new scene-quality metric. Appendix B gives the detailed nearest-work boundary and broader references, while Appendix C defines the configuration-card fields.

# 3 Experimental Setup

## 3.1 Audit design

We ask whether common evaluator choices change conclusions when the generated scene itself is frozen. For prompt $p$ and generator $m$, the generated structured scene is $s = m(p)$. A renderer produces an image $x = \mathcal{R}(s, c)$, an evaluation caption $t = \tau(p')$ wraps either the original prompt or a paraphrase, and evaluator $e$ returns $\sigma_e(x, t)$. We call the full measurement protocol $\pi = (\mathcal{R}, c, \tau, p', e, \text{preprocessing}, \text{score scale})$ and write $q_e(m, p; \pi)$ for the score of the fixed scene $m(p)$. This notation makes the audited object explicit: the scene is fixed, while the protocol that turns that scene into a number is varied.

The audit has two parts. A **configuration check** changes one evaluator-side choice at a time while holding the scene and all other choices at the default $c_0$, following the reliability side of measurement-theoretic audits (Zhao et al., 2024; Ross et al., 2024; Saxon et al., 2024). A **directional degradation check** edits the structured scene in a targeted way and asks whether scores drop on average, giving intervention-specific sensitivity evidence. The first part measures protocol sensitivity, while the second tests response to specific interventions rather than general validity. Neither is sufficient alone, since a stable score can track the wrong signal and response to one degradation need not establish broad validity. The one-factor-at-a-time design keeps the render count linear and estimates each factor conditional on the default settings.

## 3.2 Scenes, protocols, and evaluators

We use $N = 50$ prompts drawn with a fixed stratified seed from SceneEval, InstructScene, Holodeck, and Scenethesis (Tam et al., 2026; Lin & Mu, 2024; Yang et al., 2024; Li et al., 2026). The sample weights SceneEval more heavily because its object annotations match the content probes, and the resulting intervals are diagnostic prompt-bootstrap summaries rather than population estimates. Six generators produce the frozen scenes: Scenethesis, Holodeck, HSM, LayoutVLM, I-Design, and Reason-3D (Li et al., 2026; Yang et al., 2024; Pun et al., 2026; Sun et al., 2025; Çelen et al., 2024; Berdoz et al., 2026). These systems span constraint-IR synthesis, LLM procedural construction, hierarchical motifs, differentiable VLM optimization, agentic design, and reasoning-model planning. All emit structured object-level scenes with asset meshes, poses, and boxes, giving 300 fixed scenes for the audit and enabling object-level perturbations. Methods that accept a configurable text backbone use the recorded identifier `gemini-3-flash-preview` at temperature 0.7. Each generated scene is saved once and reused throughout the audit.

Every scene is rendered through one Blender Cycles pipeline. Transparent-film rendering lets us composite backgrounds after lighting, which decouples the background factor from illumination. Auto-framing keeps the scene box fixed in the image as focal length changes, and the dollhouse wall convention keeps indoor content visible under oblique views. We sweep eight evaluator-configuration factors independently: resolution, focal length, lighting, background, pitch, yaw, caption template, and paraphrase (Table 11). The default $c_0$ is a 512 px render, 50 mm focal length, city environment map, white background, top-down pitch, default yaw, the default caption template,

and the original prompt. The shared sweep has 9 resolution levels, 7 focal lengths, 8 lighting maps, 10 backgrounds, 7 pitch values, 8 yaw values, and 337 caption templates. Paraphrases are scored only at the default render because their levels are prompt-specific strings. Yaw is swept at a fixed $45^\circ$ pitch because yaw at the default top-down pitch is mostly an in-plane rotation. The camera sweep also defines a best-view test by taking the largest score over the tested poses. These factors are protocol choices, not all strict nuisance transformations, because captions can alter textual assertions and camera or resolution settings can change visible evidence. Table 11 gives the full level inventory, so the main text can focus on what each family of choices estimates.

Caption templates are all source-to-sink paths through a small grammar, giving 337 wrappers including the default *"a top-down view of {prompt}"* and the raw prompt (Appendix E). Paraphrases use the frozen file of 10 intended meaning-preserving rewrites per prompt. Duplicate caption strings collapse in scoring, giving 8–11 distinct captions per prompt. Two hosted auditors accept only about half of the 500 rewrites as meaning preserving, so paraphrase dispersion is treated as a stress test of the full text pipeline rather than a primary invariance claim (Appendix J).

We audit 19 text-scene alignment evaluators plus one scoped perceptual-quality control (Table 12). The set covers CLIP-family encoders used in text-to-3D scene papers, SigLIP, BLIP-2, BLIP-VQA, VQAScore, and a hosted judge. The CLIP-family sweep varies architecture, training data, patch size, and native resolution, which helps separate a single-checkpoint failure from a broader pattern in metrics used by the field. The hosted-judge row uses `gpt-4o-2024-11-20`. The scoring client records request and parsing exceptions as zero scores. Appendix I bounds the effect of stored zero scores with a re-scoring analysis and a deterministic sensitivity range. CLIP-IQA is a prompt-free image-quality control and is excluded from alignment-only headline counts.

### 3.3 Content probes

The content probes construct a paired stress-test scene $s^-$ and render it at $c_0$ against the original prompt. Across prompts that specify the targeted property, a responsive evaluator should tend to score the original at least as high as the edited scene. Because some generated originals may already be wrong and some prompts are generic, this is a dataset-level diagnostic expectation rather than pairwise human ground truth. Directional discrimination credits a strict drop as 1, an exact tie as 0.5, and a rise as 0, so chance is $50\%$ under symmetric random ordering.

The four single-shot probes are *keep-half*, *biggest-only*, *scrambled*, and *worst-objects*. *Keep-half* removes a random half of objects, and *biggest-only* keeps only the largest object. *Scrambled* relocates and reorients every object while preserving the object set, making it the cleanest layout probe. *Worst-objects* replaces each asset with the lowest-ranked retrieval candidate under the generator's own asset score while preserving label, pose, size, and count. Worst-objects is least clean because CLIP-family evaluators may partly recover the retrieval signal used to select the bad asset. For that reason, the results emphasize scrambling as the necessary-condition probe for spatial arrangement sensitivity.

### 3.4 Statistics and uncertainty

For a factor $f$ with levels $\ell$, let $\sigma_{s,\ell}$ be the score of scene $s$ at level $\ell$, $\bar{\sigma}_s$ its mean over levels, and $\mathbb{E}_s$ the scene average. **Dispersion** measures score movement as $D^{(f)} = \mathbb{E}_s[(\max_\ell \sigma_{s,\ell} - \min_\ell \sigma_{s,\ell})/\bar{\sigma}_s]$. **Configuration-to-signal ratio** compares within-scene configuration variance with between-generator mean-score variance at the default, $\rho^{(f)} = \mathbb{E}_s[\mathrm{Var}_\ell\, \sigma_{s,\ell}] \,/\, \mathrm{Var}_m(\mathbb{E}_p \sigma_{m,p,c_0})$. The denominator is the variance over the six generator means after averaging prompts at $c_0$. Thus $\rho > 1$ means configuration levels contribute more variance than generator means on that evaluator's own scale. It is not a statement that a configuration is more important than generator quality in a human sense. We report template $\rho$ and peak $\rho$ per evaluator. **Rank stability** uses Kendall's coefficient of concordance $W$ over factor levels, where $W = 1$ means the same order at every level. A same-scene point-estimate winner change occurs when two levels of a factor have different unique top generators while every $m(p)$ is fixed. **Best-view inflation** is the median over scenes of the largest tested-view score divided by the default-view score, restricted to positive baselines. Together these statistics separate score stability, decision robustness, and targeted probe sensitivity, which a single leaderboard number would collapse.

Table 1: Decision-level results from 50 prompts, six generators, and 300 fixed scenes, scored by 19 alignment evaluators and one render-only control. Configuration sweeps change one factor at a time from a default, so these results do not measure factor interactions.

| Audit question | Result | Scope and interpretation |
|---|---|---|
| Does configuration variance exceed generator separation? | 17/19 | Peak $\rho > 1$ for the six-generator field. |
| Does this persist under prompt resampling? | 11/19 | Bootstrap lower bound for peak $\rho$ exceeds 1. |
| Does the point-estimate winner change? | 18/19 | Across tested factor levels, without claiming a confirmed reversal. |
| Is an opposite pairwise ordering confirmed after search? | 0/19 | Evaluators with a selected reversal under a band over 110,295 contrasts. |
| Which pairwise orders survive all tested protocols? | 4–11/15 | Strict orders per alignment evaluator over a finite set of 387 protocols. |
| Does scrambling lower the score? | $\leq 67\%$ | Tie-adjusted ceiling across evaluators, including the render-only control. |

Generator comparisons use the same fixed scenes. For evaluator $e$, let $\Pi_e$ contain the baseline and all shared-level protocols, excluding scene-specific paraphrases. For generators $m, n$, define $L_e(m;\pi) = N^{-1}\sum_p q_e(m,p;\pi)$ and $\Delta_{e,mn}(\pi) = L_e(m;\pi) - L_e(n;\pi)$. If $u_{e,m}(\pi) = L_e(m;\pi) - L_e(m;\pi_0)$, then

$$\Delta_{e,mn}(\pi) = \Delta_{e,mn}(\pi_0) + u_{e,m}(\pi) - u_{e,n}(\pi). \tag{1}$$

Common protocol shifts cancel, so high dispersion or high $\rho$ is not by itself a changed comparison. We summarize the finite audit set by the protocol margin envelope

$$I_{e,mn}(\Pi_e) = \left[\min_{\pi\in\Pi_e} \Delta_{e,mn}(\pi),\ \max_{\pi\in\Pi_e} \Delta_{e,mn}(\pi)\right]. \tag{2}$$

A strictly positive or negative envelope preserves the pair ordering across tested protocols, while a sign crossing records configuration dependence. This envelope is a descriptive range, not a confidence interval, and it makes no claim about untested factor combinations, new prompt samples, or human scene quality.

Prompt bootstrap gives pointwise descriptive intervals for dispersion, $\rho$, directional discrimination, and selected pairwise margins. The bootstrap resamples the 50 prompt blocks and keeps the six generators, evaluators, and configurations jointly aligned within each prompt. For winner reversals, we also use an approximate simultaneous bootstrap band over the full searched family of 19 alignment evaluators, 15 generator pairs, and 387 shared configurations (Appendix H). This family contains 110,295 contrasts and excludes the image-only control and prompt-specific paraphrase levels. Exact aggregate ties remain ties. A point-estimate top-1 change alone carries no significance claim, and even opposite pairwise intervals do not prove either generator beats all five alternatives. Paraphrases are excluded from the rank-stability and winner-reversal search because their levels are prompt-specific strings rather than shared protocols.

## 4 Results

The scenes stay fixed throughout the audit, yet scores and point-estimate winners depend on how those scenes are presented to an evaluator. Table 1 separates score movement, generator comparisons, and targeted content sensitivity. Figure 2 shows the factor-by-evaluator dispersion behind the score-movement result. Table 2 shows how the decision measures vary across all evaluators, including the render-only quality control. Appendix K gives the complete per-factor results.

### 4.1 Render-side sensitivity: the camera moves the score

Among render factors, camera pitch and lighting produce the largest default-conditioned shifts (Figure 2). Changing pitch from top-down to eye level moves BLIP-2 ITM by 179% and the CLIP encoders by 18–33%, consistent with their documented top-down blind spot (Hou et al., 2025). Across

Table 2: Per-evaluator audit across four distinct questions. Template and peak $\rho$ compare configuration variance with between-generator variance in the six-generator field. Rank stability gives the minimum Kendall's $W$ and maximum number of distinct point-estimate winners over factors. Content accuracy averages four directional probes and is not a human-validity score. Best view is the median score inflation from selecting the highest-scoring tested camera. $^{\dagger}$ marks a per-evaluator prompt-bootstrap lower bound above one for peak $\rho$ in 11 alignment rows and the prompt-free CLIP-IQA control (Appendix K).

| | $\rho$ | | Rank stability | | Content | Best |
|---|---|---|---|---|---|---|
| **Evaluator** | template | peak | min $W$ | #top-1 | **acc. (%)** | **view** |
| CLIP ViT-B/32 | 2.05 | 3.04$^{\dagger}$ | 0.75 | 2 | 67 | 1.17 × |
| CLIP ViT-B/16 | 2.30 | 2.30$^{\dagger}$ | 0.64 | 3 | 68 | 1.14 × |
| CLIP ViT-L/14 | 3.04 | 3.04$^{\dagger}$ | 0.70 | 2 | 71 | 1.14 × |
| CLIP ViT-L/14@336 | 1.85 | 1.85$^{\dagger}$ | 0.77 | 2 | 72 | 1.12 × |
| OpenCLIP ViT-L/14 | 1.08 | 1.08 | 0.75 | 2 | 70 | 1.10 × |
| OpenCLIP ViT-H/14 | 2.04 | 2.04$^{\dagger}$ | 0.72 | 3 | 75 | 1.05 × |
| OpenCLIP ViT-bigG/14 | 1.98 | 1.98$^{\dagger}$ | 0.84 | 2 | 73 | 1.06 × |
| MetaCLIP ViT-L/14 | 1.19 | 1.19 | 0.79 | 2 | 69 | 1.11 × |
| DFN ViT-L/14 | 0.81 | 0.81 | 0.95 | 2 | 75 | 1.06 × |
| EVA-CLIP ViT-L/14 | 0.98 | 0.98 | 0.81 | 3 | 69 | 1.06 × |
| CLIPA ViT-L/14 | 2.44 | 2.44$^{\dagger}$ | 0.91 | 2 | 68 | 1.03 × |
| Long-CLIP ViT-L/14 | 1.38 | 1.38 | 0.83 | 2 | 75 | 1.06 × |
| SigLIP ViT-L/16 | 1.02 | 1.02 | 0.84 | 2 | 73 | 1.12 × |
| SigLIP 2 ViT-L/16 | 1.90 | 1.90$^{\dagger}$ | 0.87 | 2 | 74 | 1.06 × |
| BLIP-2 ITM | 2.17 | 2.17$^{\dagger}$ | 0.78 | 2 | 73 | 1.08 × |
| BLIP-2 ITC | 2.38 | 2.38$^{\dagger}$ | 0.79 | 2 | 71 | 1.06 × |
| BLIP-VQA | 0.70 | 1.58$^{\dagger}$ | 0.83 | 3 | 72 | 1.21 × |
| VQAScore | 0.80 | 1.48 | 0.87 | 3 | 71 | 1.37 × |
| GPT-4o judge | 0.39 | 1.06 | 0.94 | 1 | 74 | 1.27 × |
| CLIP-IQA | – | 2.15$^{\dagger}$ | 0.36 | 3 | 69 | 1.32 × |

eight lighting maps spanning about 5.25 exposure stops, brightness correlates $r = 0.83$ with the mean score shift (Table 8). This association does not isolate brightness from other map properties. Resolution and focal length move CLIP-family scores by roughly 7–21% each, while background is mildest at about 10%. The background result is partly mechanical, since compositing behind a transparent render leaves 13% of top-down scenes byte-identical. Lighting, focal length, and pitch can also change what a human can see, so their entire effect cannot be labeled evaluator noise.

### 4.2 Prompt-side sensitivity: the template is the largest knob

For 18 of 19 alignment evaluators, the choice among 337 caption-template grammar paths moves the score more than any render factor in the default-conditioned sweep (Figure 2). BLIP-VQA is the exception, with pitch dispersion of 42% against template dispersion of 36%. For OpenAI CLIP ViT-L/14, mean per-scene normalized range is 51% across the grammar and 28% even across common caption forms such as the bare prompt and *a top-down view of*. The scene and task prompt remain fixed, although wrappers can assert a viewpoint or image type that may be inaccurate. The hosted GPT-4o judge has template ranges of 390% and 83%, and Appendix I documents its zero-score handling and sensitivity bounds. Paraphrase results are secondary because the rewrite audit found meaning drift in about half of the rewrites (Appendix J). Template choice is also directional: the best grammar path beats the default *a top-down view of* by a median 18% (Appendix E.1). The pattern spans multiple CLIP variants: only DFN and EVA-CLIP fall below peak $\rho = 1$, while template dispersion and weak scramble responses persist in both (Tables 13 and 17). Table 3 decomposes the wording effect by matched grammar slots. The advantage of *rendering* over *photo* may partly reflect a truthful image-type description, so these shifts are protocol effects rather than uniformly evaluator defects.

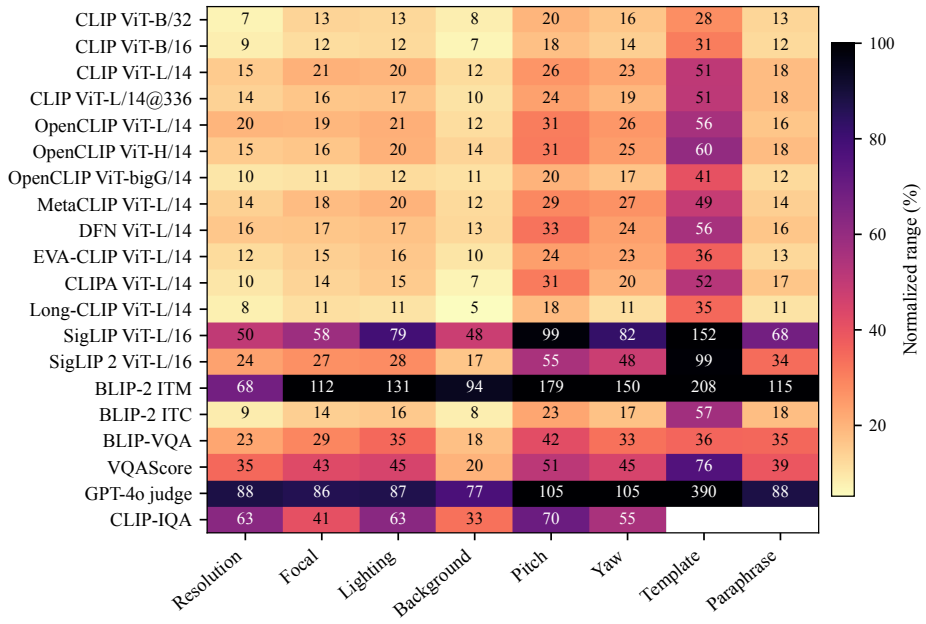



Figure 2: Score dispersion (normalized range, %) across evaluators and configuration factors in default-conditioned one-factor-at-a-time sweeps. Cells show the true values; colors are clipped at 100%. Caption template is the largest factor for 18 of 19 alignment evaluators, while pitch and lighting dominate the render factors. The prompt-free CLIP-IQA control has no caption measurements.

Table 3: Per-slot caption-template effects over the 19 text-swept evaluators. Dispersion is the range of choice means as a percentage of the slot mean, and shifts are relative to that mean. Each slot is measured on a matched subset of captions to limit confounding by co-occurring slots. *(none)* omits an optional slot.

| **Slot** | **Disp.** | **Choices** |
|---|---|---|
| Viewpoint | 19% | *top-down* +7%, *bird-eye* +5%, (none) −12% |
| Head noun | 11% | *rendering* +6%, *view* +1%, *image* −1%, *picture* −2%, *photo* −4% |
| Scene frame | 14% | *a virtual scene:* +9%, *a scene:* −2%, *scene:* −2%, (none) −5% |
| Connective | 3% | *depicting* +1%, *about* 0%, *of* 0% |
| Article | 4% | *a* +1%, (none) 0%, *an* −1% |

### 4.3 Score movement does not by itself establish a winner reversal

Peak $\rho$ divides configuration variance by the variance of six generator means, so its size depends on both the evaluator and this generator field. The ratio exceeds one for 17/19 alignment evaluators as a point estimate and retains a prompt-bootstrap lower bound above one for 11/19 (Table 15). Leaving out each generator in turn preserves the threshold for 13 alignment evaluators, but does not cover a wholly different generator field. A common score shift cancels in pairwise margins, so $\rho > 1$ alone cannot show a comparison change (Equation 1).

Aggregate rankings remain fairly concordant, with median minimum Kendall's $W = 0.80$ across all 20 evaluators, but 18 of 19 alignment evaluators have more than one top-1 generator across factor levels. At baseline, a median of 11 of 15 generator pairs are separable by paired prompt bootstrap, yet this does not confirm a reversal for any particular pair. Under the same 50 prompt blocks, 4 evaluators have opposite pointwise 95% intervals at two levels of one factor. None has opposite intervals under the approximate simultaneous band over 110,295 searched contrasts (Appendix H). The pointwise count changes from four to three under a second bootstrap seed, while the simultaneous count stays zero. We therefore report changes in the point-estimate winner, not statistically confirmed reversals after full-search correction. The absence of confirmation is not evidence that configurations have no effect. For a selected example, CLIP ViT-L/14@336 scores Scenethesis minus Reason-3D at $+1.53$ under $15°$ pitch and $-1.39$ under $90°$ pitch. Their pointwise 95% intervals, $[+0.52, +2.62]$

Table 4: Selected same-pair point-estimate reversals at two configurations. Gaps have the same first-method-minus-second-method orientation within each row, but their scales differ across evaluators. Pointwise 95% intervals exclude zero in opposite directions. Every full-search simultaneous interval includes zero. The BLIP-VQA pointwise result changes under a second bootstrap seed.

| Evaluator / factor | Method pair | Levels | Gap [pointwise 95%] | Simultaneous 95% |
|---|---|---|---|---|
| CLIP ViT-B/32 | | city | $+0.91$ $[+0.12, +1.72]$ | $[-0.88, +2.70]$ |
| Lighting | Scenethesis − I-Design | night | $-0.91$ $[-1.62, -0.20]$ | $[-2.56, +0.73]$ |
| CLIP ViT-L/14@336 | | 15° | $+1.53$ $[+0.52, +2.62]$ | $[-0.79, +3.85]$ |
| Pitch | Scenethesis − Reason-3D | 90° | $-1.39$ $[-2.57, -0.23]$ | $[-4.07, +1.29]$ |
| MetaCLIP ViT-L/14 | | 0° | $+2.31$ $[+1.03, +3.67]$ | $[-0.63, +5.26]$ |
| Pitch | Scenethesis − LayoutVLM | 75° | $-1.57$ $[-2.88, -0.22]$ | $[-4.54, +1.41]$ |
| BLIP-VQA | | T1 | $+4.33$ $[+0.13, +8.93]$ | $[-5.63, +14.29]$ |
| Template | Scenethesis − Holodeck | T2 | $-7.52$ $[-11.81, -3.67]$ | $[-17.01, +1.97]$ |

T1: "A top-down rendering about a virtual scene: {prompt}". T2: the unwrapped prompt.

and $[-2.57, -0.23]$, exclude zero in opposite directions, while both full-search simultaneous intervals include zero (Table 4). This example illustrates a configuration-dependent sample comparison, not a confirmed population winner reversal.

The margin envelopes provide a complementary descriptive result: 4–11 of 15 generator pairs per alignment evaluator retain a strict ordering across the baseline and 386 shared factor levels (Table 9). The rest cross zero somewhere in this finite protocol set. Only the hosted judge retains one winner throughout. Stable orderings are neither simultaneous confidence statements nor evidence of human-judged correctness.

### 4.4 Content probes: weak targeted arrangement sensitivity

Figure 3 compares the four targeted content probes, with per-evaluator values in Table 17. Removal probes are easier than arrangement: keeping only the largest object is detected at 74–90% tie-adjusted directional discrimination. For layout scrambling, no evaluator exceeds 67% and most score 54–62% against the intact scene. The ceiling belongs to CLIP-IQA, a control that never reads the prompt and can react to visual plausibility rather than text-layout agreement. For the hosted GPT-4o judge, 23 of 300 intact–scrambled pairs contain a zero score. Its tie-adjusted discrimination is 54.2%, and assigning arbitrary credit to those zero-score pairs yields sensitivity bounds of $[49.7, 57.3]\%$ (Appendix I). Scrambling is a dataset-level necessary-condition probe under the expectation that random relocation and reorientation should worsen spatial alignment, not a human-validated measure of layout quality.

### 4.5 Best-view selection can inflate a reported score

Selecting the highest-scoring tested pitch/yaw leaves the scene fixed but raises median scores by 1.03–1.17× for CLIP encoders, 1.37× for VQAScore, and 1.27× for the GPT-4o judge (Table 2). Those figures measure selection over saved poses and are lower bounds on the maximum score available from a larger pose search. The preferred poses are systematic: telephoto, bright, and shallow-pitch levels score highest for nearly every evaluator (Table 18).

## 5 Discussion

**Reliability is not validity.** Score stability, ranking agreement, and targeted sensitivity answer different questions. A high concordance over six generators can coexist with an unstable point-estimate winner when the leading methods are close. Our simultaneous analysis does not establish reversed generator superiority, and the scrambling probe does not replace human judgments of scene quality. Likewise, a one-signed protocol margin envelope establishes consistency of the sampled ordering across declared configurations, not correctness of that ordering. The useful outcome is scoped evidence: which direction is stable, which depends on the protocol, and which remains statistically unresolved.

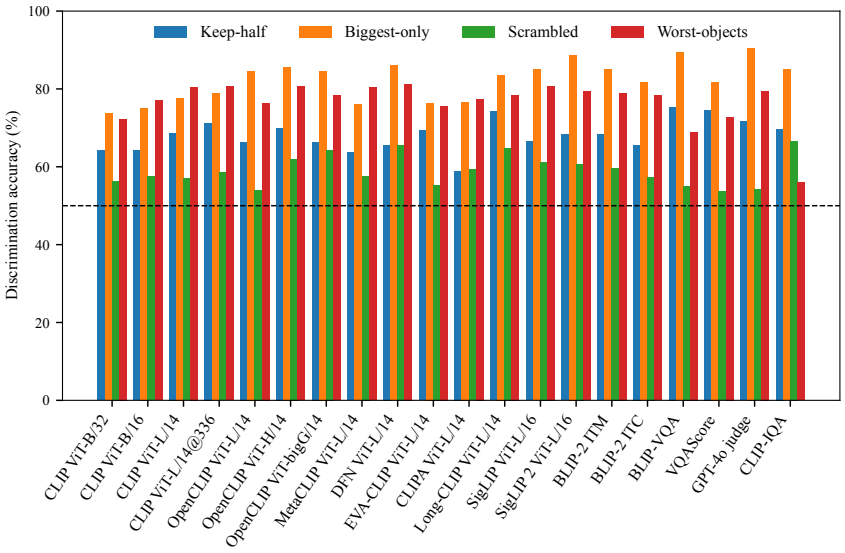



Figure 3: Tie-adjusted directional discrimination against each intact scene, with the $50\%$ chance line dashed. Removing objects is easier to detect than scrambling their layout across this fixed dataset. The scramble ceiling belongs to the prompt-free CLIP-IQA control.

**Fixed artifacts still yield different evidence.** Changing the camera can reveal a relation, hide an object, or make a scene easier to inspect. An alignment evaluator should respond when visible evidence changes, so identical scores across views are not the target. The risk is drawing a scene-level conclusion from an unspecified or selectively chosen view. Human judgments of camera-pair legibility and intact–scrambled alignment are therefore separate validation targets. Caption wrappers also need semantic audit because a correct image-type assertion and an incorrect viewpoint assertion are different controls.

**Report a contract before claiming superiority.** Reporting *(generator, score, card ID)* makes camera, caption, uncertainty, companion checks, and evidence status auditable. The proposed hash-bound provenance manifest is designed to tie numerical tables to the source revision and raw evaluator files behind them. Given six generators, $N = 50$ prompts, and an OFAT design, the current audit should be read as a diagnostic and reporting procedure. A stronger superiority claim needs an independent prompt sample and blinded human comparisons.

**How to use the audit.** First, fix the scene artifacts and identify the full evaluator protocol, including view aggregation and caption rules. Second, declare the plausible configuration set before inspecting outcomes and report pairwise margins over that set alongside the default score. A margin envelope that changes sign calls for a configuration-dependent comparison. Third, preserve prompt pairing and account for the family of comparisons searched. An interval containing zero leaves the corresponding superiority claim unresolved under that analysis, even if its point estimate is largest. Finally, report a property-matched or human-grounded sensitivity check, and distinguish request failures from observed low scores.

## 6 Conclusion

Text-to-3D evaluation is not specified by naming a metric. Across 300 frozen scenes, configuration variance often exceeds between-generator variance, and descriptive winners can change even when aggregate rankings remain relatively stable. Those changes do not survive the full-search simultaneous analysis, and weak scramble discrimination shows that ranking stability alone is not evidence of arrangement sensitivity. Pairwise protocol margin envelopes add the constructive result: some sampled comparisons keep their direction across the declared audit, while others require configuration-dependent claims. A credible comparison should report the evaluator configuration, uncertainty, stable comparison scope, and independent sensitivity evidence, with *(generator, score, card ID)* linking each number to an auditable trail.

### AI USE STATEMENT

Generative AI tools are part of the experimental pipeline: they produce the intended semantics-preserving paraphrase probes, serve as optional language backbones for applicable scene generators, and provide the hosted MLLM evaluator described in Section 3. We also used generative AI tools for code and manuscript editing. The authors take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

This work audits automatic evaluators and the reported audit introduces no human participants or personally identifiable data. Any added blinded human comparison will require the applicable ethics determination before submission. All prompts are drawn from publicly accessible sources. The generated scenes are produced by existing systems and are used only as fixed inputs to the evaluators under study. Our findings concern the reliability of evaluation metrics and are intended to improve measurement practice. We see no direct pathway to harmful use. The paper's claims concern evaluation protocols, not generator superiority: method-pair differences illustrate measurement sensitivity rather than establish which generator is best.

### REPRODUCIBILITY STATEMENT

Regenerating the reported tables requires 20 raw evaluator files with 6,000 evaluator-scene records covering the same 300 scenes, plus structured descriptions of each intact scene and its four variants. The repository provides validators for evaluator identity, scene coverage, prompt correspondence, factor levels, and deduplicated caption sets. Its hash-bound intake-manifest contract is designed to bind the source revision and raw inputs to regenerated numbers. Table 12 records evaluator identifiers, Table 11 records factor levels, and Appendix E specifies the caption grammar. For GPT scrambling, the same 300 score pairs yield 124 strict drops, 77 exact ties, and 99 rises, explaining the difference between a $41.3\%$ strict-drop rate and $54.2\%$ tie-adjusted discrimination. The hosted records identify the model behind each hosted-judge score. Freezing the saved scores fixes the analyzed evidence behind every reported number. The working repository contains the evaluation-harness source, but the complete raw records, verified intake manifest, observed evaluator card, and anonymous release archive are not present in this checkout. The reported tables therefore cannot yet be independently regenerated from this checkout alone. Zero-score cells are handled by the deterministic sensitivity range of Appendix I and by the re-scoring analysis reported there.

## REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 29304–29320, 2021.

Ali Asaria, Tony Salomone, and Deep Gandhi. A cross-model VLM-judge protocol for single-image 3D mesh quality (and why cheap proxies fall short). *arXiv preprint arXiv:2606.18451*, 2026.

Tongyuan Bai, Wangyuanfan Bai, Dong Chen, Tieru Wu, Manyi Li, and Rui Ma. FreeScene: Mixed Graph Diffusion for 3D Scene Synthesis from Free Prompts . In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5893–5903, Los Alamitos, CA, USA, June 2025. IEEE. doi: 10.1109/CVPR52734.2025.00553.

Frédéric Berdoz, Luca A. Lanzendörfer, Nick Tuninga, and Roger Wattenhofer. Text-to-scene with large reasoning models. In *AAAI Conference on Artificial Intelligence*, volume 40, pp. 2435–2443. AAAI Press, 2026. ISBN 978-1-57735-906-7. doi: 10.1609/aaai.v40i4.37229.

Zixuan Bian, Ruohan Ren, Yue Yang, and Chris Callison-Burch. HOLODECK 2.0: Vision-language-guided 3D world generation with editing. *arXiv preprint arXiv:2508.05899*, 2025.

Florian Bordes, Candace Ross, Justine T. Kao, Evangelia Spiliopoulou, and Adina Williams. Eval factsheets: A structured framework for documenting AI evaluations. *arXiv preprint arXiv:2512.04062*, 2025.

Jirayu Burapacheep, Ishan Gaur, Agam Bhatia, and Tristan Thrush. ColorSwap: A color and word order dataset for multimodal evaluation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1716–1726, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.99.

Ata Çelen, Guo Han, Konrad Schindler, Luc Van Gool, Iro Armeni, Anton Obukhov, and Xi Wang. I-Design: Personalized LLM interior designer. In *European Conference on Computer Vision (ECCV) Workshops*, volume 15624 of *Lecture Notes in Computer Science*, pp. 217–234, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-92386-9. doi: 10.1007/978-3-031-92387-6_17.

Angeline Charles, Srikant Panda, Amit Agarwal, Hitesh Laxmichand Patel, Priyaranjan Pattnayak, Bhargava Kumar, and Tejaswini Kumar. A diagnostic framework for auditing reference-free vision-language metrics. In Kentaro Inui, Sakriani Sakti, Haofen Wang, Derek F. Wong, Pushpak Bhattacharyya, Biplab Banerjee, Asif Ekbal, Tanmoy Chakraborty, and Dhirendra Pratap Singh (eds.), *International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL)*, pp. 2633–2644, Mumbai, India, December 2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. ISBN 979-8-89176-298-5. doi: 10.18653/v1/2025.ijcnlp-long.142.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829, 2023. doi: 10.1109/CVPR52729.2023.00276.

Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *International Conference on Learning Representations (ICLR)*, 2024.

Wei Deng, Mengshi Qi, and Huadong Ma. Global-Local Tree Search in VLMs for 3D Indoor Scene Generation . In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8975–8984, June 2025. doi: 10.1109/CVPR52734.2025.00839.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. WorldScore: A unified evaluation benchmark for world generation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27713–27724, October 2025. doi: 10.1109/ICCV51701.2025.02573.

Shivam Duggal, Yushi Hu, Oscar Michel, Aniruddha Kembhavi, William T. Freeman, Noah A. Smith, Ranjay Krishna, Antonio Torralba, Ali Farhadi, and Wei-Chiu Ma. Eval3D: Interpretable and fine-grained evaluation for 3D generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13326–13336, 2025. doi: 10.1109/CVPR52734.2025.01244.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. In *International Conference on Learning Representations (ICLR)*, 2024.

Avijit Ghosh, Anka Reuel, Jenny Chim, Wm. Matthew Kennedy, Srishti Yadav, Jennifer Mickel, Yanan Long, Andrew Tran, et al. Evaluation cards: An interpretive layer for AI evaluation reporting. *arXiv preprint arXiv:2606.09809*, 2026.

Zeqi Gu, Yin Cui, Zhaoshuo Li, Fangyin Wei, Yunhao Ge, Jinwei Gu, Ming-Yu Liu, Abe Davis, and Yifan Ding. ArtiScene: Language-driven artistic 3D scene generation through image intermediary. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2891–2901, 2025. doi: 10.1109/CVPR52734.2025.00275.

Yuze He, Yushi Bai, Matthieu Lin, Wang Zhao, Yubin Hu, Jenny Sheng, Ran Yi, Juanzi Li, and Yong-Jin Liu. $T^3$Bench: Benchmarking current progress in text-to-3d generation. *arXiv preprint arXiv:2310.02977*, 2023.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.595.

Kaiyuan Hou, Minghui Zhao, Lilin Xu, Yuang Fan, and Xiaofan Jiang. TDBench: A benchmark for top-down image understanding with reliability analysis of vision-language models. *arXiv preprint arXiv:2504.03748*, 2025.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, volume 36, pp. 31096–31116, 2023. doi: 10.52202/075280-1355.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20349–20360, 2023. doi: 10.1109/ICCV51070.2023.01866.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 78723–78747, 2023. doi: 10.52202/075280-3443.

Rui Huang, Guangyao Zhai, Zuria Bauer, Marc Pollefeys, Federico Tombari, Leonidas Guibas, Gao Huang, and Francis Engelmann. Video perception models for 3d scene synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 112687–112715, 2025. doi: 10.52202/085713-3763.

Nanshan Jia, Zhenyu Zhao, Sui Huang, Jingshen Wang, and Zeyu Zheng. DB-3DME: From dataset to benchmark for human-aligned automatic 3D mesh evaluation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 4631–4640, 2026.

Minseok Kang, Dongwook Choi, Gyeom Hwangbo, Seungwon Lim, Kai Tzu-iunn Ong, and Jinyoung Yeo. Apples on the table? evaluating text-guided 3D scene synthesis via fine-grained constraint verification. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2026.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023a.

Shuqing Li, Anson Y. Lam, Yun Peng, Wenxuan Wang, and Michael R. Lyu. 3D software synthesis driven by constraint-expressive intermediate representation. In *IEEE/ACM International Conference on Software Engineering (ICSE)*, ICSE '26, pp. 1354–1366, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400720253. doi: 10.1145/3744916.3787812.

Xianhang Li, Zeyu Wang, and Cihang Xie. An inverse scaling law for CLIP training. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 49068–49087, 2023b. doi: 10.52202/075280-2132.

Chenguo Lin and Yadong Mu. InstructScene: Instruction-driven 3D indoor scene synthesis with semantic graph prior. In *International Conference on Learning Representations (ICLR)*, 2024.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision (ECCV)*, volume 15067 of *Lecture Notes in Computer Science*, pp. 366–384, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-72672-9. doi: 10.1007/978-3-031-72673-6_20.

Mateusz Michalkiewicz, Sheena Bai, Mahsa Baktashmotlagh, Varun Jampani, and Guha Balakrishnan. Not all views are created equal: Analyzing viewpoint instabilities in vision foundation models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9113–9123, 2025. doi: 10.1109/ICCV51701.2025.00852.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching CLIP to count to ten. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3147–3157, 2023. doi: 10.1109/ICCV51070.2023.00294.

Hou In Derek Pun, Hou In Ivan Tam, Austin T Wang, Xiaoliang Huo, Angel X Chang, and Manolis Savva. HSM: Hierarchical scene motifs for multi-scale indoor scene generation. In *International Conference on 3D Vision (3DV)*, pp. 1356–1367. IEEE, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, volume 139, pp. 8748–8763. PMLR, 2021.

Candace Ross, Melissa Hall, Adriana Romero Soriano, and Adina Williams. What makes a good metric? evaluating automatic metrics for text-to-image consistency. In *Conference on Language Modeling (COLM)*, 2024.

Michael J. Ryan, Yanzhe Zhang, Amol Salunkhe, Yi Chu, Di Xu, and Diyi Yang. AutoMetrics: Approximate human judgments with automatically generated evaluators. In *International Conference on Learning Representations (ICLR)*, 2026.

Michael Saxon, Fatima Jahara, Mahsa Khoshnoodi, Yujie Lu, Aditya Sharma, and William Yang Wang. Who evaluates the evaluations? objectively scoring text-to-image prompt coherence metrics with T2IScoreScore (TS2). In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 85630–85657, 2024. doi: 10.52202/079017-2719.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 55565–55581, 2023. doi: 10.52202/075280-2425.

Kathakoli Sengupta, Kai Ao, and Paola Cascante-Bonilla. SceneCritic: A symbolic evaluator for 3D indoor scene synthesis. *arXiv preprint arXiv:2604.13035*, 2026. Poster at the 6th OpenSUN3D Workshop, CVPR 2026.

Anna Sokol, Elizabeth Daly, Michael Hind, David Piorkowski, Xiangliang Zhang, Nuno Moniz, and Nitesh V. Chawla. BenchmarkCards: Standardized documentation for large language model benchmarks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 91271–91289, 2025. doi: 10.52202/085713-2748.

Fan-Yun Sun, Weiyu Liu, Siyi Gu, Dylan Lim, Goutam Bhat, Federico Tombari, Manling Li, Nick Haber, and Jiajun Wu. LayoutVLM: Differentiable Optimization of 3D Layout via Vision-Language Models . In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29469–29478, Los Alamitos, CA, USA, June 2025. IEEE. doi: 10.1109/CVPR52734.2025.02744.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: Improved training techniques for CLIP at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Hou In Ivan Tam, Hou In Derek Pun, Austin T. Wang, Angel X. Chang, and Manolis Savva. SceneEval: Evaluating semantic coherence in text-conditioned 3D indoor scene synthesis. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7355–7365, 2026. doi: 10.1109/WACV61042.2026.00710.

Jiapeng Tang, Yinyu Nie, Lev Markhasin, Angela Dai, Justus Thies, and Matthias Nießner. DiffuScene: Denoising diffusion models for generative indoor scene synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20507–20518, 2024. doi: 10.1109/CVPR52733.2024.01938.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Jianyi Wang, Kelvin C.K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *AAAI Conference on Artificial Intelligence*, volume 37, pp. 2555–2563. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i2.25353.

Olivia Wiles, Chuhan Zhang, Isabela Albuquerque, Ivana Kajić, Su Wang, Emanuele Bugliarello, Yasumasa Onoe, Pinelopi Papalampidi, Ira Ktena, Christopher Knutsen, et al. Revisiting text-to-image evaluation with Gecko: On metrics, prompts, and human rating. In *International Conference on Learning Representations (ICLR)*, 2025.

Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas Guibas, Dahua Lin, and Gordon Wetzstein. GPT-4V(ision) is a human-aligned evaluator for text-to-3D generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22227–22238, 2024. doi: 10.1109/CVPR52733.2024.02098.

Hu Xu, Saining Xie, Xiaoqing Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying CLIP data. In *International Conference on Learning Representations (ICLR)*, 2024.

Yandan Yang, Baoxiong Jia, Shujie Zhang, and Siyuan Huang. SceneWeaver: All-in-one 3D scene synthesis with an extensible and self-reflective agent. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 155685–155717, 2025. doi: 10.52202/085713-4689.

Yue Yang, Fan-Yun Sun, Luca Weihs, Eli Vanderbilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, Chris Callison-Burch, Mark Yatskar, Aniruddha Kembhavi, and Christopher Clark. Holodeck: Language guided generation of 3d embodied ai environments. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16277–16287, 2024. doi: 10.1109/CVPR52733.2024.01536.

Zhaoda Ye, Xinhan Zheng, Yang Liu, and Yuxin Peng. RelScene: A benchmark and baseline for spatial relations in text-driven 3D scene generation. In *ACM International Conference on Multimedia (ACM MM)*, pp. 10563–10571, 2024. doi: 10.1145/3664647.3681653.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations (ICLR)*, 2023.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952, 2023. doi: 10.1109/ICCV51070.2023.01100.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-CLIP: Unlocking the long-text capability of CLIP. In *European Conference on Computer Vision (ECCV)*, volume 15109 of *Lecture Notes in Computer Science*, pp. 310–325, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-72982-9. doi: 10.1007/978-3-031-72983-6_18.

Yuhan Zhang, Mengchen Zhang, Tong Wu, Tengfei Wang, Gordon Wetzstein, Dahua Lin, and Ziwei Liu. 3DGen-Bench: Comprehensive benchmark suite for 3D generative models. *arXiv preprint arXiv:2503.21745*, 2025a.

Yujie Zhang, Bingyang Cui, Qi Yang, Zhu Li, and Yiling Xu. Benchmarking and learning multi-dimensional quality evaluator for text-to-3d generation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18563–18574, 2025b. doi: 10.1109/ICCV51701.2025.01725.

Dora Zhao, Jerone T. A. Andrews, Orestis Papakyriakopoulos, and Alice Xiang. Position: Measure dataset diversity, don't just claim it. In *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 60644–60673. PMLR, 2024.

Zhenyu Zhao, Nanshan Jia, Jihyeon Je, Yifu Tang, Alvin Chan, Michael Spedden, Michael V. Palleschi, Sui Huang, Jingshen Wang, and Zeyu Zheng. 3D-DefectBench: A controlled factorial study of vision-language model evaluation pipelines for fine-grained 3d generation defects. *arXiv preprint arXiv:2607.10826*, 2026.

# A APPENDIX MAP

This appendix keeps the main text focused on the audit logic and decision-level evidence. Appendix B separates our audit from the closest adjacent work. Appendix C states the configuration-card fields and scope limits. Appendices D–J provide render examples, caption grammar details, lighting scalars, protocol envelopes, selection-aware inference, hosted-score provenance, and paraphrase audit details. Appendix K gives the per-evaluator tables.

# B CLOSEST-WORK BOUNDARY

Table 5 separates a new evaluator from an audit of the measurement pipeline. "Not studied" is limited to the cited paper and does not claim absence across the full literature.

Table 5: Closest-work boundary. The cited work improves or meta-evaluates text, 2D, and 3D metrics, while the table distinguishes adjacent components from our combined intervention. "Related" marks an adjacent design rather than the same intervention.

| Work | Artifact held fixed | Render factors isolated | Caption wording isolated | Decision stability | Controlled degradation | Reporting artifact |
|---|---|---|---|---|---|---|
| AutoMetrics (Ryan et al., 2026) | Equivalent-quality outputs, not one frozen artifact | Not applicable to 3D rendering | Output rephrasing, not a 3D caption wrapper | Metric validity, not generator-winner stability | LLM-generated worse-quality outputs | Metric Cards, not a 3D configuration card |
| T2IScoreScore (Saxon et al., 2024) | Related 2D image families under a fixed prompt | Not applicable to 3D rendering | Semantic error graphs, not caption wrappers | Metric ordering, not generator-winner stability | Semantic error graphs | No analogous card in the stated contribution |
| 3D-DefectBench (Zhao et al., 2026) | Fixed assets across 84 inference designs | Camera protocol and visual input crossed | Prompt schema crossed | Factor effects and judge comparisons for defect detection | Nine human-labeled defect categories | Released data and metadata, not a reporting card |
| Cross-Model VLM-Judge (Asaria et al., 2026) | Fixed mesh pairs across 24 rendered views | Fixed multi-view rig, not render-factor isolation | Fixed judge prompt, not a caption-wording audit | Pairwise asset preference, not generator-winner stability | Face-drop degradation, not room-layout scrambling | Reproducible judge protocol, not a configuration card |
| SceneCritic (Sengupta et al., 2026) | Yes across views and repeated VLM calls | Viewpoint and repeated calls, not 8 OFAT factors | Prompt sensitivity motivated, not experimentally swept | View-dependent method-ranking reversals for one VLM judge | No fixed-artifact degradation audit | SceneOnto and a symbolic critic, not a configuration card |
| T3Bench and Eval3D (He et al., 2023; Duggal et al., 2025) | Generated assets, not a same-artifact audit | Many views, not a render-factor audit | Caption or QA pipelines, not a wording audit | Leaderboards, not stability under configurations | Janus or inconsistency probes, not room-layout scrambling | No analogous card in the stated contribution |
| VIPScene / FPVScore (Huang et al., 2025) | Generated scenes under fixed view rigs | Top-down versus first-person, not an OFAT protocol-sensitivity audit | Prompt-design ablation, not a caption-grammar sweep | Repeated-query and inter-model ranking agreement | System ablations, not paired layout edits | No analogous card in the stated contribution |
| SceneEval and LEGO-Eval (Tam et al., 2026; Kang et al., 2026) | Instruction-scene pairs, not a same-artifact configuration audit | Render or tool evidence, not a factor audit | Structured constraints, not a wording audit | Method results, not stability under configurations | Misalignment or constraint tests, not controlled pairs | No analogous card in the stated contribution |
| **This audit** | **Yes: 300 frozen scenes** | **Resolution, focal length, lighting, background, pitch, and yaw** | **Template and paraphrase** | **Ranks and top-1 under every factor** | **Object removal and layout scrambling** | **Evaluator Configuration Card** |

## B.1 BROADER RELATED WORK

Procedural scene generation includes Holodeck, Holodeck 2.0, HSM, and GLTreeSearch (Yang et al., 2024; Bian et al., 2025; Pun et al., 2026; Deng et al., 2025). Diffusion and instruction-driven alternatives include DiffuScene, InstructScene, FreeScene, and ArtiScene (Tang et al., 2024;

Lin & Mu, 2024; Bai et al., 2025; Gu et al., 2025). Constraint programs, VLM optimization, and reasoning-driven construction broaden the generator space (Li et al., 2026; Sun et al., 2025; Çelen et al., 2024; Yang et al., 2025; Berdoz et al., 2026). Their evaluation choices also vary: top-down CLIPScore, perspective or isometric renders, MLLM judges, and symbolic relation checks appear across these systems.

Image-text audits have documented failures in attribute binding, unusual viewpoints, prompt wording, and compositional relations (Hsieh et al., 2023; Burapacheep et al., 2024; Michalkiewicz et al., 2025; Huang et al., 2023). Question-answering and unified VLM metrics seek more targeted evidence (Cho et al., 2024; Huang et al., 2023; Wiles et al., 2025; Saxon et al., 2024). Object and mesh benchmarks use different units of comparison, including generated assets, pairwise asset preferences, or geometry defects (Zhang et al., 2025b; Duggal et al., 2025; Wu et al., 2024; Zhang et al., 2025a; Jia et al., 2026). Scene-oriented benchmarks add relational, geometric, first-person, and trajectory evidence (Ye et al., 2024; Kang et al., 2026; Huang et al., 2025; Duan et al., 2025). These studies motivate broader validation, while our main-text comparison is restricted to protocol effects on fixed structured room scenes.

## C Evaluator Configuration Card and scope details

Table 6 turns the reporting-artifact column of Table 5 into the fields needed to audit a single-image text-to-3D evaluator report. This domain-specific run-side contract complements BenchmarkCards, Eval Factsheets, and Evaluation Cards rather than claiming the broader card concept (Sokol et al., 2025; Bordes et al., 2025; Ghosh et al., 2026).

Table 6: Paper-ready Evaluator Configuration Card checklist. A complete JSON card records these fields, validates against schema version `1.0`, and receives a deterministic `ecc-v1-sha256:...` identifier.

| Card field | Required report content |
|---|---|
| Schema and status | `card_type`, `schema_version`, evidence status, and the validator-produced card identifier. |
| Scene artifact source | Dataset, generator set, scene and prompt counts, frozen-scene rule, and artifact provenance. |
| Render protocol | Renderer, camera pose, focal length, resolution, lighting, background, random seeds, and post-processing. |
| Caption protocol | Prompt template, paraphrase rule, prompt source, and whether wording was frozen before scoring. |
| View policy | Fixed, averaged, or selected view, declared views, and selection rule. |
| Evaluator identity | Model or metric name, checkpoint/API snapshot, instruction, decoding settings, and call date. |
| Score uncertainty | Interval or bootstrap rule, resampling unit, comparison family, decision thresholds, multiplicity policy, and whether claims cover scores, ranks, or top methods. |
| Rank stability | Rank range or winner changes across declared configurations. |
| Companion check | Geometry-grounded, multi-view, human, or explicit `not_reported` companion validation. |
| Released or missing evidence | Frozen scenes, render grids, raw score files, evaluator responses, hash manifest, and any known missing evidence. |

The audit is intrinsic and does not estimate which score difference a person would consider meaningful. Its configuration sweeps keep the scene and task prompt fixed, but caption wrappers and render settings can change the evaluator's input semantics or visible evidence. The OFAT design does not estimate interactions, and external validity still requires blinded human comparison of representative intact, scrambled, and camera-varied scenes. The judge results use one GPT-4o configuration and one rating instruction, so they do not isolate model capability or establish generalization to other MLLMs. The rank and $\rho$ results are properties of this six-generator field and an $N = 50$ sample, while every effect magnitude remains specific to the Blender pipeline and prompt sample.

## D QUALITATIVE RENDERS

Figure 4 shows the same generated scene under two render-side configuration factors, making concrete that these are ordinary choices rather than exotic manipulations.

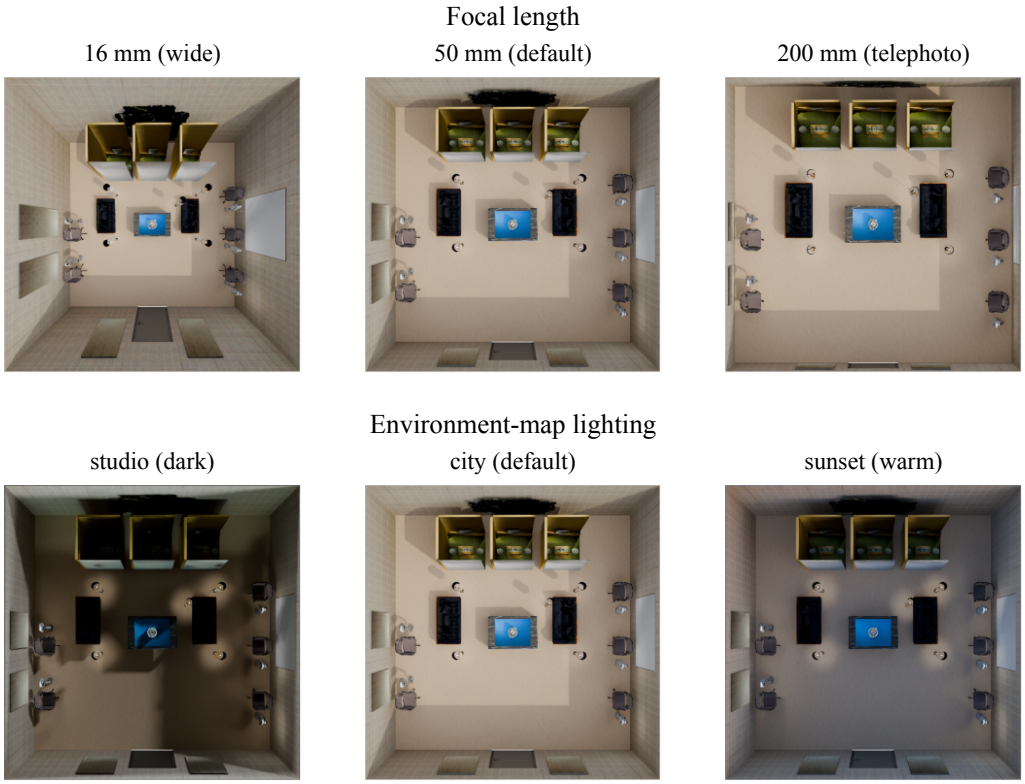



Figure 4: The same scene under two configuration factors. Top: focal length 16/50/200 mm at a fixed, auto-framed top-down view. Bottom: environment-map lighting (studio/city/sunset). Figure 1 shows the pitch sweep in the main paper.

## E TEMPLATE GRAMMAR

The 337 evaluation templates of Section 3.2 are the source-to-sink paths of a small grammar. Table 7 reads the grammar as an ordered sequence of slots (the form used to build a caption by hand), and Figure 5 draws the same grammar as a directed acyclic token graph (the form that makes the path count explicit). The token nodes are written in lower case, but the first letter of every realized caption is capitalized, so the default template reads *"A top-down view of* {*prompt*}*"* rather than *"a top-down . . . "*. The one exception is the degenerate identity path, whose only token is the prompt, which is emitted verbatim.

### E.1 WHICH WORDING THE EVALUATORS REWARD

Section 4.2 reports how far the template moves the score. Here we ask which choices in the grammar are responsible, by parsing all 337 captions back into their slots and averaging the score of each slot choice over the text-swept evaluators. Table 3 gives, per slot, its dispersion and each choice's mean shift from the slot's own average. To keep a slot from being confounded by the company it keeps, each is measured on a matched subset. Head nouns are compared only where a viewpoint word is present, so all five nouns span the same caption skeletons, and connectives only where a head noun is present. The head-noun choice is not neutral. *Rendering* scores highest and *photo* lowest, a range of

Table 7: The template grammar as ordered slots. A caption picks at most one entry from each slot in order (subject to the adjacency in Figure 5) and always ends with the prompt. The connective is present exactly when a head noun is. The default template *"a top-down view of* {*prompt*}*"* takes the article, the *top-down* viewpoint, *view*, the connective *of*, no scene frame, and the prompt.

| Slot | Status | Choices |
|---|---|---|
| Article | *optional* | a, an |
| Viewpoint | *optional* | top-down, bird-eye |
| Head noun | *optional* | image, photo, picture, rendering, view |
| Connective | *with a noun* | of, about, depicting |
| Scene frame | *optional* | scene:, a scene:, a virtual scene: |
| Prompt | *required* | {prompt} |

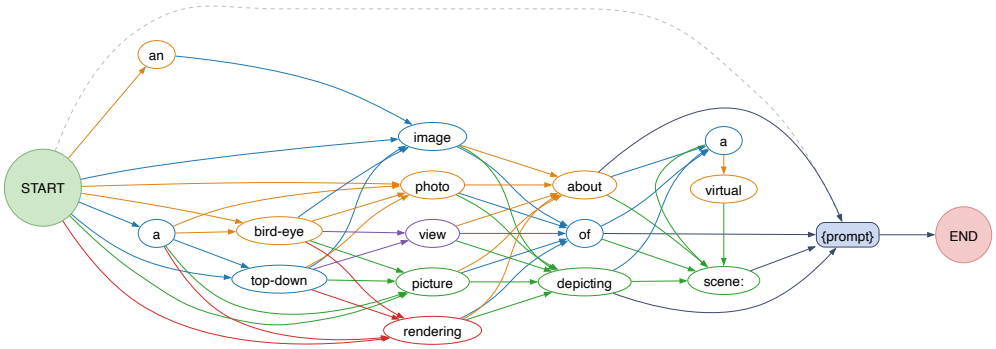



Figure 5: The template grammar as a directed acyclic token graph. Each of the 337 templates is exactly one source-to-sink path (START $\rightarrow \cdots \rightarrow$ {prompt} $\rightarrow$ END). Each edge is colored by the node it points into, and every node is outlined in its own color, so an edge and its target always match. The dashed edge from START to {prompt} is the degenerate identity template.

11% of the slot mean, and rendering leads for all 19 evaluators. Because these images are renders, part of this gap could reflect appropriate image-type matching rather than an evaluator defect. It still changes the reported scale without changing the artifact, so matching the caption convention is necessary for a numerical comparison. The scene frame matters at least as much: naming *a virtual scene:* sits about 9% above the slot average and omitting it about 5% below, a range of 14%. Naming a viewpoint at all is worth more than which viewpoint, since *top-down* beats *bird-eye* by only a few percent while dropping the viewpoint word costs about 19% relative to *top-down*, the slot's full range. The article and the connective are close to inert. The article deserves a note, because its raw marginal is misleading. *An* occurs only before the vowel-initial *image* and never beside a viewpoint word, so every *An image* caption also forgoes the viewpoint, and lumping them together makes *an* look like a large penalty. Measured among viewpoint-free captions, where *a*, *an*, and no article are genuine alternatives, all three fall within a few percent, and the choice between *a* and *an* is in any case fixed by the following word rather than free. These choices compound. The single best template is a *top-down rendering of a virtual scene:* for almost every evaluator, and it beats the default *"a top-down view of"* by a median of 18% (up to 52%), with the default ranking only 202 of 337. An author who reports the same renders under the best caption rather than the default raises every alignment number without touching a pixel, which is the template analogue of the best-view inflation of Section 4.5.

## F Environment-map scalars

Table 8 lists the two scalars used to summarize the eight lighting maps in Section 4.1: a log-average exposure value (EV, computed from the map's luminance) and a correlated color temperature (CCT, from the chromaticity of the map's total radiated energy). The rightmost column is each map's mean

score shift, standardized per evaluator and averaged over evaluators and scenes, so that maps with a positive value tend to raise the score and maps with a negative value tend to depress it. Brightness (EV) correlates $r = 0.83$ with this aggregate shift, but this association does not isolate brightness from other map properties or establish the same effect for every scene.

Table 8: The eight lighting maps ordered dark to bright. EV is a log-average exposure value (lower is darker). CCT is correlated color temperature (lower is warmer). The default map "city" is the brightest. Mean $z$-shift is the map's average standardized score shift across evaluators and scenes.

| Environment map | EV | CCT (K) | Mean $z$-shift |
|---|---|---|---|
| studio | -6.41 | 7802 | -0.067 |
| night | -5.13 | 5051 | -0.173 |
| courtyard | -3.70 | 5759 | -0.017 |
| sunrise | -3.24 | 6004 | -0.044 |
| forest | -2.22 | 7513 | +0.075 |
| interior | -2.20 | 5420 | +0.077 |
| sunset | -2.02 | 8785 | +0.012 |
| city$^\star$ | -1.17 | 6444 | +0.136 |

## G PAIRWISE COMPARISON OVER THE TESTED PROTOCOLS

Table 9 applies the protocol margin envelope to every unordered generator pair. For an alignment evaluator, the audit set contains the baseline and all 386 shared levels of the seven rankable factors, with yaw measured at the documented $45°$ pitch. Repeated default-equivalent levels are retained as recorded protocols and do not alter the envelope endpoints. Scene-specific paraphrases are excluded because they are not a shared configuration level. The quality control has no caption sweep, so its 50 recorded baseline and render protocols are reported separately. The categories retain exact aggregate ties and describe the saved point estimates only.

Table 9: Pairwise ordering over the finite tested protocol set. Strict ordering means that a generator pair keeps the same nonzero margin sign throughout the set. A sign crossing includes both positive and negative margins, while a tie boundary reaches zero without crossing signs. A stable winner beats all five alternatives at every tested protocol. These descriptive ranges are not confidence intervals or statements of human-grounded superiority. The hosted judge's zero-score handling is documented in Appendix I, and the image-only control, marked *qc*, lacks a caption sweep and is summarized over its 50 render-only protocols.

| Evaluator | Protocols | Strict ordering | Sign crossing | Tie boundary | Stable winner |
|---|---|---|---|---|---|
| CLIP ViT-B/32 | 387 | 5/15 | 10/15 | 0/15 | none |
| CLIP ViT-B/16 | 387 | 4/15 | 11/15 | 0/15 | none |
| CLIP ViT-L/14 | 387 | 4/15 | 11/15 | 0/15 | none |
| CLIP ViT-L/14@336 | 387 | 5/15 | 10/15 | 0/15 | none |
| OpenCLIP ViT-L/14 | 387 | 5/15 | 10/15 | 0/15 | none |
| OpenCLIP ViT-H/14 | 387 | 5/15 | 10/15 | 0/15 | none |
| OpenCLIP ViT-bigG/14 | 387 | 5/15 | 10/15 | 0/15 | none |
| MetaCLIP ViT-L/14 | 387 | 9/15 | 6/15 | 0/15 | none |
| DFN ViT-L/14 | 387 | 11/15 | 4/15 | 0/15 | none |
| EVA-CLIP ViT-L/14 | 387 | 7/15 | 8/15 | 0/15 | none |
| CLIPA ViT-L/14 | 387 | 11/15 | 4/15 | 0/15 | none |
| Long-CLIP ViT-L/14 | 387 | 5/15 | 10/15 | 0/15 | none |
| SigLIP ViT-L/16 | 387 | 8/15 | 7/15 | 0/15 | none |
| SigLIP 2 ViT-L/16 | 387 | 10/15 | 5/15 | 0/15 | none |
| BLIP-2 ITM | 387 | 5/15 | 10/15 | 0/15 | none |
| BLIP-2 ITC | 387 | 8/15 | 7/15 | 0/15 | none |
| BLIP-VQA | 387 | 6/15 | 9/15 | 0/15 | none |
| VQAScore | 387 | 8/15 | 7/15 | 0/15 | none |
| GPT-4o judge | 387 | 11/15 | 4/15 | 0/15 | Scenethesis |
| CLIP-IQA | 50 | 0/15 | 15/15 | 0/15 | none$^{qc}$ |

## H SELECTION-AWARE SAME-PAIR INFERENCE

A changed sample winner need not imply a well-separated reversal, especially after searching many configurations. We construct a single contrast family containing all 19 alignment evaluators, 15 generator pairs, and 387 configurations: the baseline plus the 386 shared levels of the seven rankable factors. This gives $J = 110,295$ paired contrasts, excluding the image-only control and the prompt-specific paraphrase strings. For contrast $j$, let $d_{p,j}$ be the score difference between its two generators on prompt $p$, with empirical mean $\bar{d}_j$ and paired standard error $\widehat{\mathrm{SE}}_j$. We resample the 50 prompt blocks jointly across every evaluator, method, and configuration, preserving all within-prompt dependencies. Using $B = 2000$ draws with seed 42, we compute the bootstrap maximum standardized deviation and its empirical $95\%$ quantile:

$$T_b = \max_{j:\widehat{\mathrm{SE}}_j>0} \frac{|\bar{d}_j^{*(b)} - \bar{d}_j|}{\widehat{\mathrm{SE}}_j}, \qquad I_j = \left[\bar{d}_j - q_{0.95}\widehat{\mathrm{SE}}_j,\ \bar{d}_j + q_{0.95}\widehat{\mathrm{SE}}_j\right]. \tag{3}$$

The standard errors are fixed at their original-sample values, not re-estimated within each draw, and exact-constant contrasts are ineligible for inferential declarations. The resulting critical value is $q_{0.95} = 4.397$. This is approximate exploratory bootstrap coverage for the fixed six-generator field and empirical prompt distribution, not a finite-sample family-wise guarantee or a prevalence estimate for all scene generators.

For each evaluator and factor, we consider pairs that are distinct unique sample winners at two levels and ask whether their two intervals exclude zero in opposite directions. All candidate pairs and levels are already in the simultaneous family, including those subsequently selected for display. The primary seed yields 18 evaluators with descriptive reversals, 4 with opposite pointwise intervals, and 0 with opposite simultaneous intervals (Table 4). At seed 1234, the pointwise count is three because the BLIP-VQA example is marginal, while the descriptive and simultaneous counts are unchanged. The baseline-only comparison in the primary analysis yields three evaluators with opposite pointwise intervals, a different estimand from searching two arbitrary levels of the same factor. None of these pairwise comparisons establishes that either selected winner is superior to all five alternatives.

## I HOSTED-SCORE PROVENANCE AND ZERO-SCORE SENSITIVITY

The GPT-4o score file records 120,270 stage-score cells, of which 6,758 are zero. These zero cells are dependent measurements, not independent model runs. The scoring client returns zero for request or parsing exceptions, so a stored zero can represent a genuine zero rating or a failed request. We therefore re-score the zero cells among the 300 audited scramble pairs and report deterministic sensitivity bounds that cover both interpretations. Table 10 decomposes the 300 paired scramble outcomes instead of equating ties with detected degradation.

Table 10: GPT-4o scramble outcomes from the same 300 intact and scrambled score pairs. Intervals use the joint prompt bootstrap of Appendix H and are pointwise descriptive summaries. Tie-adjusted discrimination credits a drop as one, a tie as one-half, and a rise as zero.

| **Outcome** | **Count** | **Percent [95% interval]** |
|---|---|---|
| Strict drop | 124 | 41.3 [35.0, 47.7] |
| Exact tie | 77 | 25.7 [20.7, 31.0] |
| Rise | 99 | 33.0 [27.0, 38.7] |
| Tie-adjusted discrimination | | 54.2 [48.5, 59.5] |

There are 13 zero baselines, 17 zero scrambled scores, and 7 pairs with both scores zero, giving 23 pairs involving at least one zero. The remaining 277 pairs contribute 149 units of directional credit, yielding $53.8\%$ on this selected subset. That subset is not an unbiased failure-corrected estimate. Allowing arbitrary credit for every zero-involved pair instead bounds full-sample discrimination by $[149/300, 172/300] = [49.7, 57.3]\%$. These bounds capture uncertainty about the zero scores, not sampling uncertainty.

We re-score the 30 stored zero-valued cells, 13 intact-side and 17 scrambled-side, with GPT-4o at temperature 0 under a bounded retry policy that treats a zero-valued response as a suspected call

failure and repeats it for up to six attempts. Thirteen cells return a nonzero score within the retry limit with a mean resolved score of 20.0, and the remaining 17 stay zero across every attempt. Every persistent zero carries an explicit low-alignment justification in its reasoning, naming the requested objects missing from the render, and 13 of the 17 persist in HSM scenes. The retry policy resolves the unstable zeros, while the persistent ones are consistent with genuine zero ratings under this re-scoring run. The deterministic sensitivity range of Table 10 bounds the effect of the zero-involved pairs on the full-sample estimand.

## J PARAPHRASE AUDIT

The frozen prompt file stores 10 rewrites for each of the 50 prompts, and we audit every rewrite with two hosted auditors at temperature 0. The auditors are `deepseek-v4-flash`, the model that generated the rewrites, and the independent hosted judge GPT-4o. Each auditor receives the original prompt and one rewrite and answers a fixed rubric question: does the rewrite preserve every spatial and semantic requirement of the original, including objects, counts, spatial relations, room type, and style. The auditors accept 245 of 500 rewrites ($49.0\%$) and 238 of 500 ($47.6\%$), and they agree on 376 of 500 rewrites ($75.2\%$). The similarity score stored with each rewrite does not separate the two outcomes: mean similarity is 0.236 for accepted and 0.237 for rejected rewrites under one auditor, and 0.232 versus 0.240 under the other. A typical joint rejection rewrites *"a bar"* as *"a pub"* or *"a tavern"*, which changes the requested object. A machine verdict is not a human judgment.

## K PER-EVALUATOR DETAIL

Table 11: Evaluator-configuration factors, each swept independently with all others held at the default $c_0$ ($\star$). The cross-factor baseline uses a white background. The background sweep ranges over the listed gray and chromatic fills.

| Factor | # | Levels |
|---|---|---|
| Resolution (pixel) | 9 | 196, 224, 256, 336, 384, 448, 512$^\star$, 768, 1024 |
| Focal length (mm) | 7 | 16, 24, 35, 50$^\star$, 85, 100, 200 |
| Lighting (environment map) | 8 | city$^\star$, courtyard, forest, interior, night, studio, sunrise, sunset |
| Background (RGB) | 10 | 0, 65, 118, 128, 186, 204, 255$^\star$ gray, plus red, green, blue |
| Pitch (degrees) | 7 | 0$^\star$, 15, 30, 45, 60, 75, 90 |
| Yaw (degrees, at pitch $45^\circ$) | 8 | 0$^\star$, 45, 90, 135, 180, 225, 270, 315 |
| Template | 337 | source-to-sink paths of the grammar (Appendix E). Defaults to *"a top-down view of* $\{ p' \}$*"* |
| Paraphrase | 8–11 | distinct strings from the original prompt plus 10 intended semantics-preserving LLM rewrites |

Table 12: The 20 evaluators under audit, grouped by family. CLIP-IQA is reported apart from the alignment metrics because it does not condition on the prompt.

| Group | Members |
|---|---|
| CLIP (OpenAI) | ViT-B/32, ViT-B/16, ViT-L/14, ViT-L/14@336 (Radford et al., 2021) |
| CLIP (open data) | OpenCLIP ViT-L/14, ViT-H/14, ViT-bigG/14 (Cherti et al., 2023), MetaCLIP (Xu et al., 2024), DFN (Fang et al., 2024), EVA-CLIP (Sun et al., 2023), and CLIPA (Li et al., 2023b) |
| CLIP (long-text) | Long-CLIP (Zhang et al., 2024) |
| Sigmoid objective | SigLIP (Zhai et al., 2023) and SigLIP 2 (Tschannen et al., 2025) |
| BLIP-2 (Li et al., 2023a) | ITM and ITC heads |
| VQA-based | BLIP-VQA (Huang et al., 2023) and VQAScore (Lin et al., 2024) |
| MLLM judge | GPT-4o |
| Quality (scoped) | CLIP-IQA (Wang et al., 2023) |

Tables 13, 14, 16, and 17 give the per-factor numbers summarized in Table 2: full dispersion, the configuration-to-signal ratio $\rho$, rank stability, and tie-adjusted directional discrimination. Figure 6

visualizes prompt-bootstrap uncertainty for peak $\rho$. Table 15 also reports leave-one-generator-out sensitivity.

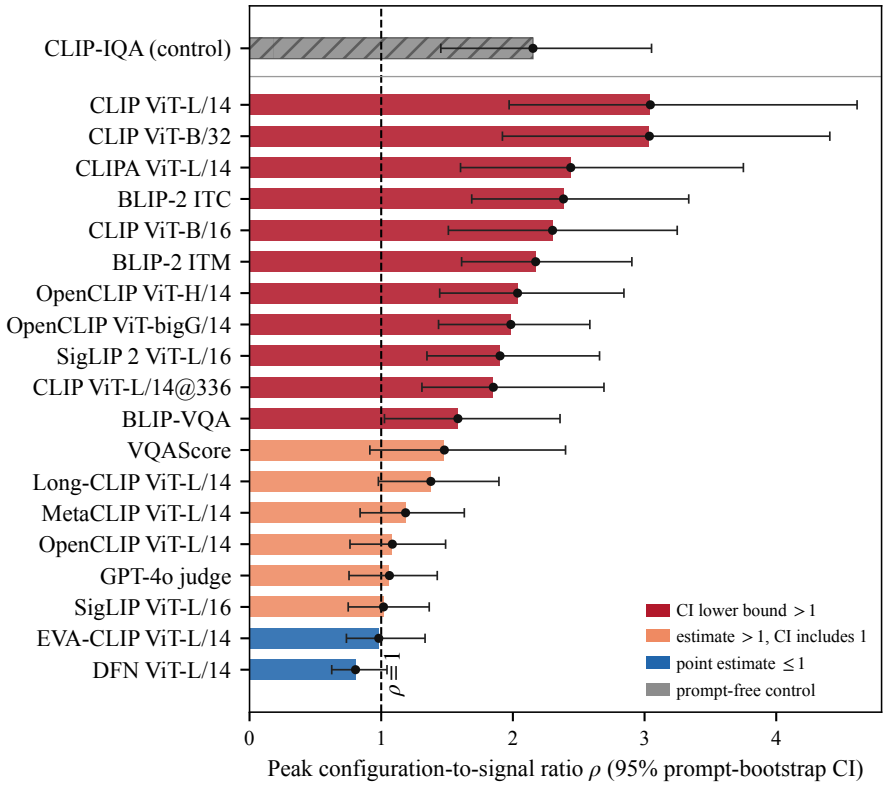



Figure 6: Peak configuration-to-signal ratio for the fixed six-generator field, with per-evaluator prompt-bootstrap 95% intervals (50 prompts with generators fixed). The lower bound exceeds one for 11 of 19 alignment evaluators. CLIP-IQA is shown separately as a prompt-free quality control. A high ratio alone does not establish a changed pairwise ordering.

**Additional result details.** Within the OpenCLIP family, moving from ViT-L/14 to ViT-bigG/14 raises tie-adjusted scramble discrimination from 54% to 64%, but also raises template $\rho$ from 1.08 to 1.98. The $\rho$ denominator uses the population variance over the six generator means. Using sample variance would rescale every $\rho$ by $5/6$ and still leaves 13 alignment evaluators and the prompt-free control above 1. The GPT-4o scramble outcomes are decomposed into strict drops, exact ties, and rises in Appendix I.

**Benchmark-source robustness.** For the hosted GPT-4o judge, template sensitivity and weak scramble discrimination also occur outside the dominant SceneEval prompts. Splitting that judge by prompt source, template dispersion is 342% on SceneEval versus 501% on the other three benchmarks, and scramble directional discrimination is 55% versus 51%. Both effects are stronger off SceneEval for this judge, but this source breakout does not establish the same pattern for all 20 evaluators.

## L Best and worst level per render factor

Table 18 lists, for each evaluator and render factor, the level with the highest and the lowest mean score over the 300 scenes, written *best/worst*. No level is established here as the human-grounded

Table 13: Score dispersion (normalized range, %) per evaluator and factor. Larger is worse. *Res.* is resolution, *Bg.* is background, *Tmpl.* is template, and *Para.* is paraphrase.

| Evaluator | Res. | Focal | Light | Bg. | Pitch | Yaw | Tmpl. | Para. |
|---|---|---|---|---|---|---|---|---|
| CLIP ViT-B/32 | 7 | 13 | 13 | 8 | 20 | 16 | 28 | 13 |
| CLIP ViT-B/16 | 9 | 12 | 12 | 7 | 18 | 14 | 31 | 12 |
| CLIP ViT-L/14 | 15 | 21 | 20 | 12 | 26 | 23 | 51 | 18 |
| CLIP ViT-L/14@336 | 14 | 16 | 17 | 10 | 24 | 19 | 51 | 18 |
| OpenCLIP ViT-L/14 | 20 | 19 | 21 | 12 | 31 | 26 | 56 | 16 |
| OpenCLIP ViT-H/14 | 15 | 16 | 20 | 14 | 31 | 25 | 60 | 18 |
| OpenCLIP ViT-bigG/14 | 10 | 11 | 12 | 11 | 20 | 17 | 41 | 12 |
| MetaCLIP ViT-L/14 | 14 | 18 | 20 | 12 | 29 | 27 | 49 | 14 |
| DFN ViT-L/14 | 16 | 17 | 17 | 13 | 33 | 24 | 56 | 16 |
| EVA-CLIP ViT-L/14 | 12 | 15 | 16 | 10 | 24 | 23 | 36 | 13 |
| CLIPA ViT-L/14 | 10 | 14 | 15 | 7 | 31 | 20 | 52 | 17 |
| Long-CLIP ViT-L/14 | 8 | 11 | 11 | 5 | 18 | 11 | 35 | 11 |
| SigLIP ViT-L/16 | 50 | 58 | 79 | 48 | 99 | 82 | 152 | 68 |
| SigLIP 2 ViT-L/16 | 24 | 27 | 28 | 17 | 55 | 48 | 99 | 34 |
| BLIP-2 ITM | 68 | 112 | 131 | 94 | 179 | 150 | 208 | 115 |
| BLIP-2 ITC | 9 | 14 | 16 | 8 | 23 | 17 | 57 | 18 |
| BLIP-VQA | 23 | 29 | 35 | 18 | 42 | 33 | 36 | 35 |
| VQAScore | 35 | 43 | 45 | 20 | 51 | 45 | 76 | 39 |
| GPT-4o judge | 88 | 86 | 87 | 77 | 105 | 105 | 390 | 88 |
| CLIP-IQA | 63 | 41 | 63 | 33 | 70 | 55 | – | – |

Table 14: Configuration-to-signal ratio $\rho$ per evaluator and factor. Values $\geq 1$ mean that within-scene configuration variance equals or exceeds the variance between generator mean scores on that evaluator's scale. *Res.* is resolution, *Bg.* is background, *Tmpl.* is template, and *Para.* is paraphrase.

| Evaluator | Res. | Focal | Light | Bg. | Pitch | Yaw | Tmpl. | Para. |
|---|---|---|---|---|---|---|---|---|
| CLIP ViT-B/32 | 0.34 | 1.34 | 1.09 | 0.46 | 3.04 | 2.01 | 2.05 | 0.97 |
| CLIP ViT-B/16 | 0.36 | 0.88 | 0.86 | 0.24 | 2.07 | 1.35 | 2.30 | 0.72 |
| CLIP ViT-L/14 | 0.57 | 1.35 | 1.05 | 0.35 | 2.10 | 1.37 | 3.04 | 0.78 |
| CLIP ViT-L/14@336 | 0.30 | 0.54 | 0.49 | 0.17 | 1.13 | 0.64 | 1.85 | 0.47 |
| OpenCLIP ViT-L/14 | 0.28 | 0.36 | 0.34 | 0.11 | 0.90 | 0.50 | 1.08 | 0.20 |
| OpenCLIP ViT-H/14 | 0.25 | 0.34 | 0.44 | 0.23 | 1.16 | 0.56 | 2.04 | 0.34 |
| OpenCLIP ViT-bigG/14 | 0.25 | 0.35 | 0.41 | 0.28 | 1.23 | 0.69 | 1.98 | 0.38 |
| MetaCLIP ViT-L/14 | 0.21 | 0.44 | 0.46 | 0.17 | 1.00 | 0.75 | 1.19 | 0.21 |
| DFN ViT-L/14 | 0.14 | 0.17 | 0.16 | 0.09 | 0.66 | 0.29 | 0.81 | 0.13 |
| EVA-CLIP ViT-L/14 | 0.21 | 0.39 | 0.42 | 0.17 | 0.93 | 0.70 | 0.98 | 0.22 |
| CLIPA ViT-L/14 | 0.18 | 0.43 | 0.43 | 0.09 | 1.80 | 0.67 | 2.44 | 0.45 |
| Long-CLIP ViT-L/14 | 0.15 | 0.43 | 0.31 | 0.07 | 1.06 | 0.39 | 1.38 | 0.30 |
| SigLIP ViT-L/16 | 0.14 | 0.29 | 0.34 | 0.10 | 0.74 | 0.36 | 1.02 | 0.31 |
| SigLIP 2 ViT-L/16 | 0.21 | 0.32 | 0.28 | 0.10 | 1.08 | 0.63 | 1.90 | 0.39 |
| BLIP-2 ITM | 0.12 | 0.42 | 0.39 | 0.12 | 1.17 | 0.53 | 2.17 | 0.43 |
| BLIP-2 ITC | 0.10 | 0.32 | 0.33 | 0.09 | 0.76 | 0.37 | 2.38 | 0.43 |
| BLIP-VQA | 0.28 | 0.55 | 0.63 | 0.15 | 1.58 | 0.98 | 0.70 | 1.07 |
| VQAScore | 0.50 | 0.85 | 0.76 | 0.14 | 1.48 | 1.02 | 0.80 | 0.55 |
| GPT-4o judge | 0.48 | 0.60 | 0.48 | 0.30 | 1.06 | 0.75 | 0.39 | 0.42 |
| CLIP-IQA | 2.15 | 0.93 | 1.44 | 0.70 | 1.94 | 1.29 | – | – |

correct setting, yet the columns are strikingly consistent. Marginally under the shared default configuration, telephoto focal lengths, bright maps, and shallow pitches score highest for nearly every evaluator, while wide-angle, dark, and eye-level score lowest. A shared and directional bias is more damaging for reporting than random noise because it creates a predictable opportunity for score selection (Section 4.5). Standardizing one configuration improves comparability only when it is declared and shared. It cannot by itself establish construct validity or repair conclusions drawn under hidden configurations.

Table 15: Uncertainty and generator-set sensitivity of the configuration-to-signal ratio, and baseline generator separability. *Peak* $\rho$ is the point estimate of Table 2. *Prompt 95% CI* resamples the $N = 50$ prompts with the six generators fixed. *Leave-one-gen range* recomputes peak $\rho$ after dropping each generator. *Top-1/2 gap* is the paired-bootstrap difference between the two baseline leaders, and $^*$ marks an interval that excludes zero. This column does not identify the actual winner pair at every configuration level. CLIP-IQA has no alignment Top-1/2 gap because it never conditions on the prompt, and its other rank columns are quality-control pseudo-rankings.

| Evaluator | Peak $\rho$ | Prompt 95% CI | Leave-one-gen range | Top-1/2 gap [95% CI] |
|---|---|---|---|---|
| CLIP ViT-B/32 | 3.04 | [1.92, 4.41] | [2.51, 9.06] | +0.91 [+0.12, +1.73]$^*$ |
| CLIP ViT-B/16 | 2.30 | [1.51, 3.25] | [1.91, 7.92] | +0.33 [-0.64, +1.28] |
| CLIP ViT-L/14 | 3.04 | [1.97, 4.61] | [2.54, 9.60] | +1.12 [-0.24, +2.38] |
| CLIP ViT-L/14@336 | 1.85 | [1.31, 2.69] | [1.52, 5.94] | +1.71 [+0.46, +2.87]$^*$ |
| OpenCLIP ViT-L/14 | 1.08 | [0.76, 1.49] | [0.89, 6.45] | +2.59 [+1.11, +4.03]$^*$ |
| OpenCLIP ViT-H/14 | 2.04 | [1.44, 2.84] | [1.70, 12.69] | +1.49 [+0.01, +2.90]$^*$ |
| OpenCLIP ViT-bigG/14 | 1.98 | [1.44, 2.58] | [1.64, 3.94] | +3.35 [+2.01, +4.72]$^*$ |
| MetaCLIP ViT-L/14 | 1.19 | [0.84, 1.63] | [1.00, 3.02] | +1.24 [+0.19, +2.27]$^*$ |
| DFN ViT-L/14 | 0.81 | [0.62, 1.04] | [0.66, 2.73] | +2.24 [+0.90, +3.53]$^*$ |
| EVA-CLIP ViT-L/14 | 0.98 | [0.74, 1.33] | [0.81, 2.69] | +0.01 [-0.78, +0.73] |
| CLIPA ViT-L/14 | 2.44 | [1.60, 3.75] | [2.06, 3.17] | +0.53 [-0.35, +1.54] |
| Long-CLIP ViT-L/14 | 1.38 | [0.98, 1.89] | [1.12, 10.51] | +0.66 [-0.33, +1.51] |
| SigLIP ViT-L/16 | 1.02 | [0.75, 1.36] | [0.83, 6.10] | +1.11 [+0.11, +2.14]$^*$ |
| SigLIP 2 ViT-L/16 | 1.90 | [1.35, 2.66] | [1.60, 6.44] | +1.24 [+0.47, +1.99]$^*$ |
| BLIP-2 ITM | 2.17 | [1.61, 2.90] | [1.75, 21.92] | +10.68 [+0.67, +20.31]$^*$ |
| BLIP-2 ITC | 2.38 | [1.69, 3.34] | [1.99, 11.50] | +1.71 [+0.01, +3.28]$^*$ |
| BLIP-VQA | 1.58 | [1.02, 2.36] | [1.25, 12.88] | +1.26 [-4.22, +6.91] |
| VQAScore | 1.48 | [0.91, 2.40] | [1.28, 4.15] | +1.44 [-4.01, +7.06] |
| GPT-4o judge | 1.06 | [0.76, 1.43] | [0.88, 2.42] | +18.55 [+10.00, +27.55]$^*$ |
| CLIP-IQA | 2.15 | [1.45, 3.05] | [1.82, 3.50] | – |

Table 16: Rank stability per factor, reported as *Kendall's W / number of distinct top-1 winners* across levels. $W = 1$ with a single winner is perfect stability. Paraphrases are excluded because their levels are per-scene strings that cannot be aligned across scenes. *Res.* is resolution, *Bg.* is background, and *Tmpl.* is template.

| Evaluator | Res. | Focal | Light | Bg. | Pitch | Yaw | Tmpl. |
|---|---|---|---|---|---|---|---|
| CLIP ViT-B/32 | 0.91/1 | 0.89/2 | 0.83/2 | 0.95/1 | 0.75/2 | 0.87/2 | 0.94/1 |
| CLIP ViT-B/16 | 0.97/1 | 0.89/1 | 0.75/3 | 0.97/2 | 0.64/2 | 0.73/1 | 0.89/2 |
| CLIP ViT-L/14 | 0.91/1 | 0.89/1 | 0.70/2 | 0.95/1 | 0.82/2 | 0.94/2 | 0.91/2 |
| CLIP ViT-L/14@336 | 0.87/1 | 0.95/1 | 0.84/2 | 0.97/1 | 0.85/2 | 0.77/1 | 0.89/1 |
| OpenCLIP ViT-L/14 | 0.85/1 | 0.94/1 | 0.75/2 | 0.93/1 | 0.80/2 | 0.89/1 | 0.97/1 |
| OpenCLIP ViT-H/14 | 0.87/1 | 0.95/1 | 0.72/2 | 0.95/1 | 0.81/3 | 0.95/1 | 0.94/1 |
| OpenCLIP ViT-bigG/14 | 0.90/1 | 0.95/1 | 0.89/1 | 0.94/1 | 0.84/2 | 0.88/1 | 0.97/1 |
| MetaCLIP ViT-L/14 | 0.95/1 | 1.00/1 | 0.95/1 | 0.99/1 | 0.79/2 | 0.90/2 | 0.97/1 |
| DFN ViT-L/14 | 0.97/1 | 1.00/1 | 0.95/1 | 1.00/1 | 0.95/2 | 0.97/1 | 1.00/1 |
| EVA-CLIP ViT-L/14 | 0.95/2 | 0.93/2 | 0.90/3 | 0.96/1 | 0.92/2 | 0.81/2 | 0.94/3 |
| CLIPA ViT-L/14 | 1.00/1 | 0.95/2 | 0.97/2 | 0.96/1 | 0.91/2 | 0.98/1 | 0.96/2 |
| Long-CLIP ViT-L/14 | 0.97/1 | 0.91/2 | 0.83/2 | 0.96/1 | 0.86/2 | 0.90/1 | 0.88/1 |
| SigLIP ViT-L/16 | 0.99/1 | 1.00/1 | 0.84/2 | 0.98/1 | 0.92/2 | 0.99/1 | 1.00/1 |
| SigLIP 2 ViT-L/16 | 0.95/1 | 0.95/1 | 0.96/1 | 0.98/1 | 0.87/2 | 1.00/1 | 0.94/1 |
| BLIP-2 ITM | 0.94/1 | 0.88/2 | 0.78/2 | 0.89/1 | 0.92/2 | 0.92/1 | 0.89/2 |
| BLIP-2 ITC | 0.94/1 | 0.97/2 | 0.79/2 | 0.93/1 | 0.91/2 | 1.00/1 | 0.88/2 |
| BLIP-VQA | 0.84/3 | 0.83/3 | 0.86/2 | 0.96/2 | 0.87/2 | 0.83/1 | 0.87/3 |
| VQAScore | 0.96/2 | 0.92/2 | 0.87/3 | 1.00/1 | 0.91/2 | 0.91/1 | 0.94/2 |
| GPT-4o judge | 0.97/1 | 0.97/1 | 0.94/1 | 0.98/1 | 1.00/1 | 0.96/1 | 0.96/1 |
| CLIP-IQA | 0.57/3 | 0.81/2 | 0.74/2 | 0.73/2 | 0.36/3 | 0.65/3 | – |

Table 17: Tie-adjusted directional discrimination (%): a strict drop receives 1, an exact tie 0.5, and a rise 0 (chance = 50%). The intended order is a dataset-level diagnostic expectation rather than per-pair human ground truth. The final column, "Mean (4)," is the plain average of the four probe columns, and is the value reported as *Content acc.* in Table 2.

| Evaluator | Keep-half | Biggest-only | Scrambled | Worst-objects | Mean (4) |
|---|---|---|---|---|---|
| CLIP ViT-B/32 | 64 | 74 | 56 | 72 | 67 |
| CLIP ViT-B/16 | 64 | 75 | 57 | 77 | 68 |
| CLIP ViT-L/14 | 69 | 78 | 57 | 80 | 71 |
| CLIP ViT-L/14@336 | 71 | 79 | 58 | 81 | 72 |
| OpenCLIP ViT-L/14 | 66 | 85 | 54 | 76 | 70 |
| OpenCLIP ViT-H/14 | 70 | 86 | 62 | 81 | 75 |
| OpenCLIP ViT-bigG/14 | 66 | 85 | 64 | 78 | 73 |
| MetaCLIP ViT-L/14 | 64 | 76 | 58 | 80 | 69 |
| DFN ViT-L/14 | 66 | 86 | 66 | 81 | 75 |
| EVA-CLIP ViT-L/14 | 69 | 76 | 55 | 76 | 69 |
| CLIPA ViT-L/14 | 59 | 77 | 59 | 77 | 68 |
| Long-CLIP ViT-L/14 | 74 | 84 | 65 | 78 | 75 |
| SigLIP ViT-L/16 | 66 | 85 | 61 | 81 | 73 |
| SigLIP 2 ViT-L/16 | 68 | 89 | 61 | 79 | 74 |
| BLIP-2 ITM | 68 | 85 | 60 | 79 | 73 |
| BLIP-2 ITC | 66 | 82 | 57 | 78 | 71 |
| BLIP-VQA | 75 | 89 | 55 | 69 | 72 |
| VQAScore | 75 | 82 | 54 | 73 | 71 |
| GPT-4o judge | 72 | 90 | 54 | 80 | 74 |
| CLIP-IQA | 70 | 85 | 67 | 56 | 69 |

Table 18: Highest- and lowest-scoring level (*best/worst*) of each render factor per evaluator, by mean score over the 300 scenes. Background is given as its gray value (0 black to 255 white) or color name. No level is established as human-grounded ground truth, yet the best entries are dominated by telephoto, bright, and shallow-pitch levels and the worst by wide-angle, dark, and eye-level.

| Evaluator | Resolution | Focal | Lighting | Background | Pitch | Yaw |
|---|---|---|---|---|---|---|
| CLIP ViT-B/32 | 224/1024 | 85/16 | interior/studio | 204/green | 45/0 | 180/45 |
| CLIP ViT-B/16 | 336/224 | 200/16 | forest/night | 255/65 | 45/0 | 180/135 |
| CLIP ViT-L/14 | 448/224 | 200/16 | city/night | 255/green | 30/90 | 90/180 |
| CLIP ViT-L/14@336 | 1024/384 | 200/16 | forest/night | 255/green | 15/90 | 45/180 |
| OpenCLIP ViT-L/14 | 1024/196 | 200/16 | city/night | 128/green | 15/90 | 0/225 |
| OpenCLIP ViT-H/14 | 768/196 | 200/16 | city/night | 128/red | 15/90 | 0/225 |
| OpenCLIP ViT-bigG/14 | 768/196 | 100/16 | city/night | 255/green | 15/90 | 0/45 |
| MetaCLIP ViT-L/14 | 224/196 | 200/16 | interior/night | 186/red | 0/90 | 0/45 |
| DFN ViT-L/14 | 768/224 | 200/16 | city/studio | 65/green | 0/90 | 0/225 |
| EVA-CLIP ViT-L/14 | 768/224 | 100/16 | city/night | 65/blue | 30/90 | 0/225 |
| CLIPA ViT-L/14 | 1024/196 | 200/16 | city/night | 186/blue | 0/90 | 90/225 |
| Long-CLIP ViT-L/14 | 448/224 | 200/16 | city/night | 65/red | 30/90 | 315/270 |
| SigLIP ViT-L/16 | 768/196 | 200/16 | city/night | 0/green | 0/90 | 0/225 |
| SigLIP 2 ViT-L/16 | 256/196 | 200/16 | city/night | 0/green | 0/90 | 0/225 |
| BLIP-2 ITM | 1024/224 | 200/16 | city/night | 255/green | 15/90 | 0/45 |
| BLIP-2 ITC | 336/196 | 200/16 | city/night | 0/green | 15/90 | 270/225 |
| BLIP-VQA | 224/336 | 200/16 | city/night | red/128 | 45/0 | 0/225 |
| VQAScore | 768/196 | 16/100 | city/night | 65/0 | 90/0 | 0/180 |
| GPT-4o judge | 1024/196 | 16/85 | forest/night | 0/255 | 15/90 | 0/90 |
| CLIP-IQA | 196/448 | 200/35 | interior/night | 0/255 | 0/15 | 180/45 |