BRIDGEGEOM: DIAGNOSING FINE-GRAINED GEOMETRIC PERCEPTION UNDER STRUCTURAL ALIASING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models often recognize a bridge as suspension, cable-stayed, or arch, but this does not show that they perceive the geometry that identifies the bridge itself. We study this gap under *structural aliasing*, where objects share a coarse engineering template and identity rests on weak local geometric cues. We introduce BridgeGeom, a diagnostic benchmark of 971 bridges, 7,555 standardized images, and 10,826 questions, whose core tasks fix the structural type so that only geometric evidence can distinguish instances. Scaling from Qwen3.5-4B to 27B lifts name-image assignment from 19.00% to 32.30%, but same-bridge verification rises only from 70.90% to 75.90% and height comparison from 55.40% to 60.30%. A structural equation over the task measurements yields a dominant scale factor that follows a scaling law with model size, and the geometric blind spot hides inside this factor. We separate it by decomposing the factor into general, image-text, and geometric capability, and measure the geometric capability at 1.057 for Qwen3.5-4B, 1.281 for 27B, and 1.396 for GPT-5.6-sol. Retrieval augmentation lifts the image-text capability but never the geometric capability, and the blind spot is thus contained within the scaling law yet separable from overall capability and shortcut use.

1 INTRODUCTION

Vision-language models can already recognize what an object is, what it is called, which category it belongs to, and how tall it is. This paper studies the next step, whether a model can distinguish different instances within the same category and read the structural geometry that makes them different once category and name lose their discriminative power. Existing questions ask for a name or a number, which can be produced from image-text associations without reading the geometry of the object itself. We refer to the remaining non-geometric information that can eliminate candidates as the *geometric margin*: the larger the margin, the easier it is for a model to avoid geometric comparison, and the less an accuracy score reveals about the evidence used.

Decision rules that ignore the object itself and rely on background or other substitute cues are known as *shortcut learning* (Geirhos et al., 2020). When tanks are photographed under blue sky and non-tanks under overcast grey, a model can answer from the background without learning the object itself (Lones, 2021), and benchmarks that admit text-only shortcuts leave this substitution undetected (Ma et al., 2026). Existing visual-question-answering and multimodal benchmarks mainly measure broad visual understanding and rarely test whether instance discrimination follows category recognition (Antol et al., 2015; Hudson & Manning, 2019; Liu et al., 2024; Li et al., 2023a; Fu et al., 2023; Yue et al., 2024). Spatial reasoning benchmarks focus on relations, layouts, and compositional logic without fixing the category as a control condition (Johnson et al., 2017; Thrush et al., 2022; Yu et al., 2025; Yang et al., 2025; Cheng et al., 2025; Xu et al., 2025; Wasi et al., 2026). Geometric reasoning benchmarks remain symbolic (Luo et al., 2026), and probing studies find that continuous geometry is present in frozen visual features but is not expressed along the text pathway (Shkolnikov, 2026). Fine-grained recognition, re-identification, and cross-view geo-localization benchmarks rely on appearance, background, place, or camera pose, rather than object-level identity under a fixed structural type (Krause et al., 2013; Maji et al., 2013; Li et al., 2023b; He et al., 2024; Lu et al., 2025b; Zuo et al., 2024; Talreja et al., 2026; Lu et al., 2025a; Wang et al., 2023; Hu et al., 2022; Garg et al., 2024). Instance-level identity has begun to be tested directly, on whether two images depict the same object (Marsili et al., 2026) or the same landmark (Kristoffersen et al., 2026). The

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.

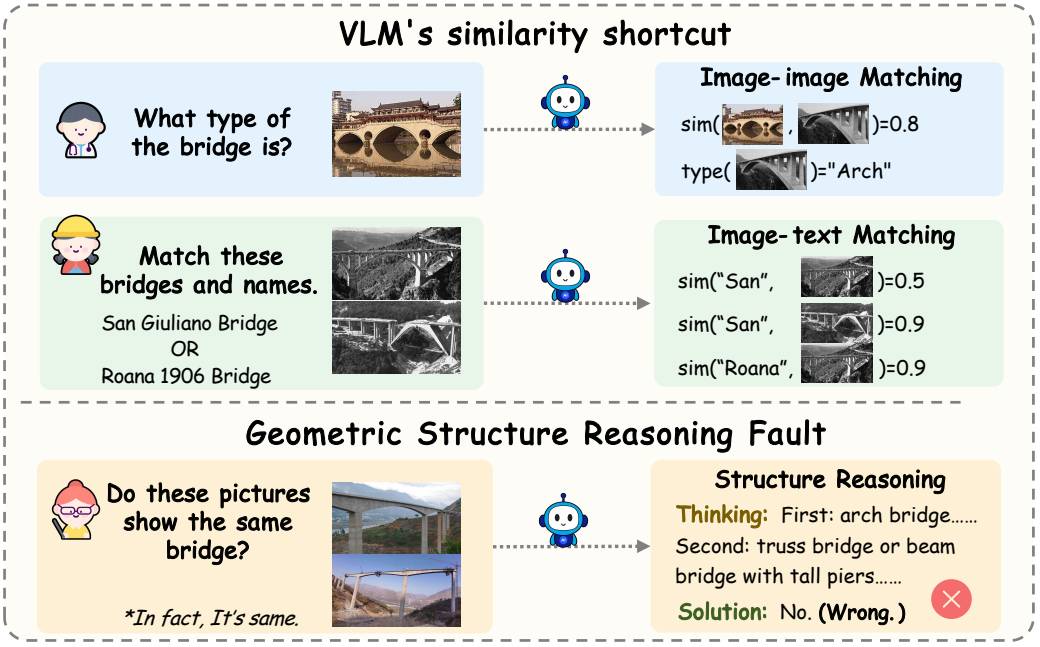

Figure 1: Two representative failure modes of current VLMs. (a) Similarity shortcut, where the model matches bridges by image and text similarity scores rather than structural geometry. (b) Geometric structure reasoning fault, where the model misidentifies the structural type of the same bridge in two views and answers incorrectly.

existing evaluation landscape therefore does not measure the contribution of structural geometry as the geometric margin decreases.

The difficulty comes from the asymmetry between the two answer routes, as illustrated in Figure 1. Names and numerical values can be produced from the text space, whereas structural geometry must be read from the image, and structural differences decay during cross-modal alignment because the text distribution cannot express them. Training targets are still names or numerical values, so no gradient requires the model to use geometry even though the encoder preserves the necessary details. What is missing is the task requirement that makes geometry determine the answer, and shortcut learning can therefore replace geometric perception while accuracy cannot separate the two routes.

We introduce BridgeGeom to study this problem on bridges, whose geometry is directly constrained by the surrounding terrain. The benchmark contains 971 bridges, 7,555 standardized images, and 10,826 questions across 8 reference categories, with core tasks that fix the structural type so that distinguishing same-type bridges relies on structural geometry alone. The seven tasks cover bridge memory, structural matching, and physical metric, with the core discrimination tasks ordered by decreasing geometric margin. We use signal detection theory to separate discrimination from response tendency, partition negative pairs by country to distinguish geometry from background cues, and combine cross-format prediction, degradation, retrieval, and scale fitting to characterize geometric ability, shortcut use, and external evidence.

The experiments quantify a geometric blind spot. Category recognition reaches 66.30% while full-group assignment reaches only 34.40%, and small models approach chance when the geometric margin is smallest. Country cues carry roughly one third of same-type discrimination, and the gap does not close with scale. Low resolution weakens geometric discrimination while preserving the shortcut cue, and greyscale does the reverse. Retrieval supplements evidence but never geometric perception. The blind spot is thus contained within the scaling law yet separable from overall capability and shortcut use.

2 Related Work

Multimodal Evaluation Benchmarks The history of multimodal evaluation is in effect a history of removing cues. The first generation asked whether models could answer questions about images at all (Antol et al., 2015), and later suites scaled this to broader understanding (Yue et al., 2024). As the suites saturated, the field discovered that high scores did not require the images, the failure mode formalized as shortcut learning (Geirhos et al., 2020). The benchmarks built in response work by prohibition and filtration. VisuLogic forbids language shortcuts outright (Xu et al., 2025), and Real-3DQA filters questions answerable from text alone, finding that 3D question answering collapses once they are removed (Ma et al., 2026). Spatial reasoning benchmarks tightened the visual evidence, from real-world relations (Yu et al., 2025; Wasi et al., 2026) to constraint-governed spaces where geometry, topology, and physical feasibility pin down the answer (Yang et al., 2026), while geometric reasoning left the image entirely and became the translation of drawing code (Luo et al., 2026). The pattern across these moves is the same. Difficulty is created by taking away the signal the model was actually using, and each removal reveals that the model was never using the signal the designer intended. Diagnosing which signal is in use remains hard, since the biases a model exploits need not coincide with annotated attributes (Vitry et al., 2026) and correction methods resist reproducible comparison (Delzer & Bender, 2026). What no generation has removed is the category itself. Every benchmark either varies the category or leaves a non-geometric signal in place. BridgeGeom completes this trajectory by fixing the structural type and removing appearance, place, and text together, so that the only remaining signal is the geometry of the structure.

Geometric Perception and Fine-Grained Recognition Research on geometric perception divides into two questions, whether the encoder contains geometric information and whether the model uses it. The first question was answered by probing. Frozen visual features carry continuous geometric measurements that the text pathway cannot express (Shkolnikov, 2026), and vision foundation models encode part-level geometric structure (Zhang et al., 2026b). The information survives only under favorable input conditions, since perception saturates below roughly twelve pixels (Li et al., 2026), loses invariance to rotation and scaling as semantic content becomes sparse (Qiu et al., 2026), and separates information loss from generalisation failure under compression (Zhang et al., 2026c), and it is not a side effect of reasoning, since shrinking the language backbone damages visual capability at least as much (Endo & Yeung-Levy, 2026). The second question was answered by benchmarks and by output tasks, and the answer is negative wherever the test permits another signal. Fine-grained benchmarks distinguish similar categories through texture and appearance (Krause et al., 2013; Li et al., 2023b) or through place and pose (Talreja et al., 2026), so success does not require structure. The instance-level tests that have begun to appear, on everyday objects (Marsili et al., 2026) and landmarks (Kristoffersen et al., 2026), find discrepancy detection far below human level (Weng et al., 2026) and attribute misbinding hidden by aggregate accuracy (Xu et al., 2026b). Output tasks fail in the same direction, with measurement reading weak (Lin et al., 2026), metric depth available only after pixel reference and camera ambiguity are removed (Cai et al., 2026; Zheng et al., 2026), millimetre tolerances unenforced in CAD generation (Gong et al., 2026; Lu et al., 2026), and inspection strong on semantics but weak on metric localization (Zhong et al., 2026; Çelik et al., 2026). No existing test connects the two answers, because none forces the model to rely on geometry when it is the only distinguishing signal. BridgeGeom provides this test by fixing the structural type and leaving local geometry as the only difference between instances.

Retrieval-Augmented Multimodal Evaluation Research on multimodal retrieval has concentrated on the retrieval interface and the retrieval unit. UniRAG provides a unified retrieval interface for large vision-language models (Sharifymoghaddam et al., 2025). EchoSight adds joint image-text re-ranking on top of visual retrieval (Yan & Xie, 2024). VisRAG indexes rendered pages as images rather than extracting their text (Shi et al., 2024). MoRE routes each query across text, image, and table experts (Peng et al., 2025). Nyx extends retrieval to mixed-modality corpora, in which queries and documents both contain text and images, and trains the retriever with feedback from the downstream VLM (Zhang et al., 2026a). KBMR studies the retrieval unit at the entity level and shows that CLIP-style dual encoders rank visually similar but distinct entities together (Xu et al., 2026a). A second direction studies how the retrieved evidence is used. Under context-memory conflict, models answer from parametric knowledge rather than the retrieved evidence (Jia et al., 2026). At scale, retrieval improves smaller VLMs and degrades larger ones (Anugraha et al., 2026). These evaluations

report end-to-end accuracy, in which a knowledge gain and a perception gain are indistinguishable. BridgeGeom fixes the images and gives bridge memory, structural matching, and physical metric the same inputs, so a gain on name recall and name-image assignment with a flat verification score traces to knowledge access, and a gain on verification itself traces to geometric perception, which is what the MRAG experiments measure.

3 THE BRIDGEGEOM BENCHMARK

3.1 DATASET CONSTRUCTION PIPELINE

BridgeGeom benchmark organises seven tasks along three diagnostic dimensions, each probing a distinct aspect of geometric perception (Figure 2; Table 1).

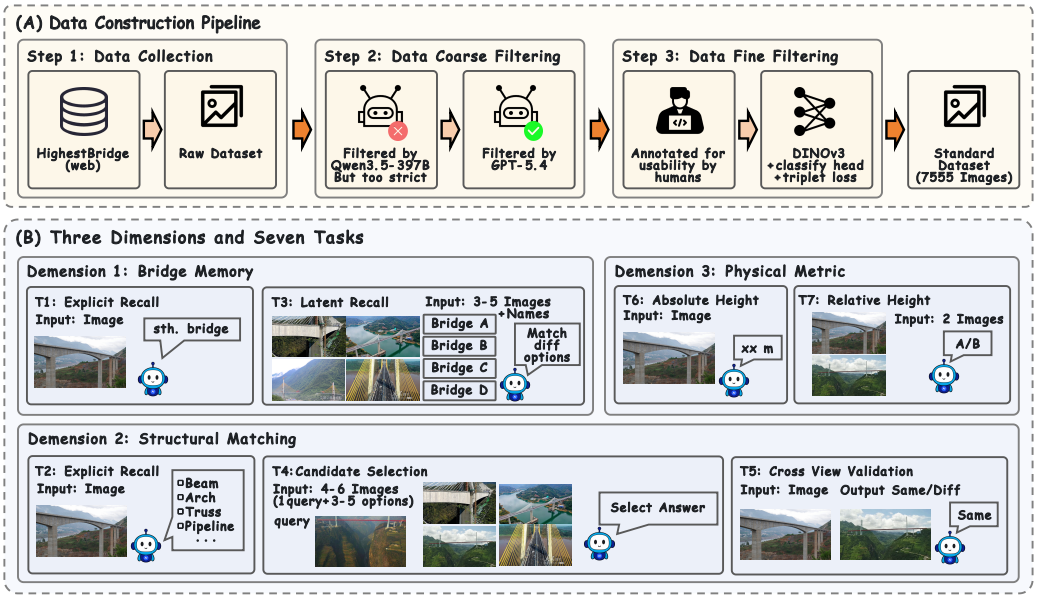

Figure 2: Overview of the BridgeGeom benchmark design.

Images come from the public HighestBridges database (HighestBridges, 2025), which records bridge names, locations, source category descriptions, and deck-to-ground heights. A three-stage filtering pipeline produces 7,555 standardised images.

Stage 1: Automated non-photographic filtering. GPT-5.4 answers a binary prompt on whether an image is a real bridge photograph. Engineering diagrams, route maps, architectural renderings, and close-up material shots are excluded. Filtering uses GPT-5.4 rather than any evaluated model, so that no model's own training corpus shapes the images it is later scored on.

Stage 2: Manual validation set. Human annotators label about 5% of the Stage-1 output as usable or unusable; unusable covers blurred, occluded, and non-bridge-centric shots. These labels serve as ground truth for Stage 3.

Stage 3: DINOv3-based filtering. The remaining 95% of Stage-1 images are encoded with DINOv3 (Oquab et al., 2025). A linear classifier trained on the Stage-2 annotations using triplet loss predicts whether each image is usable. Images classified as usable are retained; predicted outliers are discarded. Within each bridge, the retained images form a similarity graph, and we keep only the largest connected component to remove views that fail to cluster with the bridge's other photographs. All retained files are resized to 1920×1080 JPEG.

Source category descriptions lack a unified taxonomy, requiring an operational mapping. GPT-5.4 assigns each bridge to one of eight reference categories in five independent runs; the majority label is retained, with vote consistency recorded and unresolved cases reviewed manually. The eight labels are: Beam, Arch, Cable-stayed, Suspension, Truss, and Cantilever plus Footbridge and Pipeline.

The resulting corpus averages 7.78 images per bridge (median 6, range 5-27), yielding 4,855 T1 questions; all seven tasks total 10,826 questions and 19,826 image inputs. Category composition is shown in Figure 3. Country metadata covers 971 bridges across 32 countries, of which 736 (75.80%) are in China, reflecting the source database's concentration on high-clearance structures.

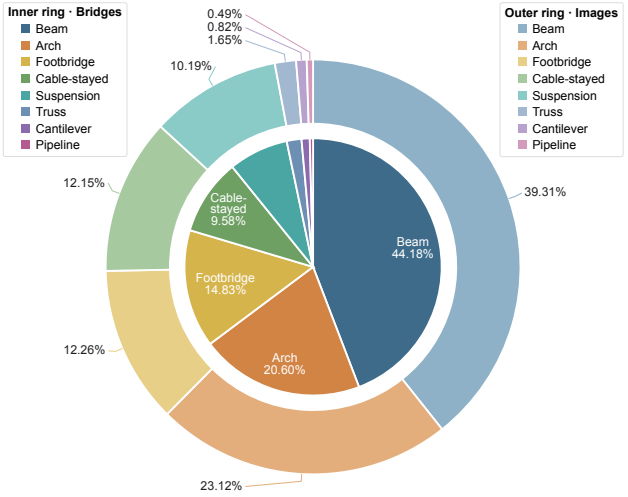

Figure 3: Dataset composition by reference category.

3.2 Data Analyze

Bridge Memory (T1, T3). T1 asks the model to name a bridge from a single image. T3 presents three to five same-type bridge images with the same number of candidate names and asks the model to match them.

Table 1: Seven tasks along three diagnostic dimensions.

Dimension	Task	Items	Metric
Bridge memory	T1	4,855	acc
Structural match	T2	1,000	acc
Bridge memory	T3	1,000	item/group acc
Structural match	T4	1,000	acc
Structural match	T5	1,000	acc
Physical metric	T6	971	MAE/MRE
Physical metric	T7	1,000	acc

Structural Matching (T2, T4, T5). T2 assigns a bridge image to one of eight reference categories. T4 presents a query image with four shuffled options, including one image of the same bridge, images from one same-type distractor bridge, and optional cross-type fillers. The fillers preserve a category-based elimination cue, while the same-type distractor requires fine-grained visual comparison. T5 asks whether two images show the same bridge, with negative pairs drawn from the same structural type and no category-based elimination. Together, T4 and T5 measure instance discrimination under two closely related comparison formats.

Physical Metric (T6, T7). T6 asks the model to estimate a bridge's deck height from a single image. T7 presents two bridge images and asks which bridge is taller. Together, they test whether visual appearance supports absolute and relative physical-scale estimation.

3.3 Evaluation Metrics

T1, T2, T4, T5, and T7 are scored by exact match, and we report accuracy against the chance level of each task, which is $1/971$ for T1, $1/8$ for T2, 25% for T4, and 50% for T5 and T7.

T3 is scored at two granularities. Item accuracy averages the proportion of correct name-image pairs within each group over the 1,000 groups, so that groups of different sizes contribute equally. Group accuracy counts a group as correct only when every assignment in it is correct. An assignment that is missing, repeated, or outside the candidate list is scored as incorrect at the item level and makes its group incorrect. Chance levels are 26.39% per item and 7.55% per group, the expectation of a

random permutation. The two scores are reported together because they bracket the same behaviour, with item accuracy crediting partial matching that group accuracy rejects.

T6 reports the mean absolute error and the mean relative error,

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{h}_i - h_i|, \qquad \mathrm{MRE} = \frac{100}{N}\sum_{i=1}^{N}\frac{|\hat{h}_i - h_i|}{h_i}, \tag{1}$$

where $\hat{h}_i$ is the predicted deck height in metres, h_i the recorded height, and $N = 971$. Lower is better for both, and T6 has no uniform chance baseline, since a constant guess produces a different error on every bridge.

4 Experiments

4.1 Models and Protocol

We evaluate ten models: Qwen3.5 dense checkpoints at 0.8B, 2B, 4B, 9B, and 27B, the Qwen3.5-35B-A3B MoE variant (Qwen Team, 2026), Gemma 4 E2B and E4B (Gemma Team et al., 2026), and the closed-source GPT-5.5 and GPT-5.6-sol models (OpenAI, 2026a;b). All models share identical frozen questions, prompts, and parsers, with greedy decoding and a 2,048-token thinking budget. MRAG and degradation experiments use Qwen3.5-4B to separate retrieval and image-signal effects from model scale.

The experiments comprise three groups. The scale series examines how model capacity affects the recognition-to-discrimination gap. The MRAG study tests whether retrieved evidence improves recognition and structural matching. The degradation study separates the effects of color and spatial resolution on visual discrimination.

The MRAG study shares an index of text records, attribute tables, knowledge cards, and bridge images for all 971 bridges, excluding the query image from retrieval. UniRAG prepends retrieved image-text pairs (Sharifymoghaddam et al., 2025), EchoSight supplies text through two-stage ranking (Yan & Xie, 2024), VisRAG retrieves rendered pages in PDF and PNG (Shi et al., 2024), and MoRE routes queries across text, image, and table experts (Peng et al., 2025).

4.2 Main Results

Table 2 reports the performance of all ten models across the seven tasks, together with five retrieval configurations on a fixed Qwen3.5-4B backbone.

Bridge memory. Exact bridge-name recall remains below 1% for all open-source models. When candidate bridge names are provided, Qwen3.5 item-level T3 accuracy increases from 25.65% at 0.8B to 57.03% at 35B-A3B, while full-group accuracy increases from 3.60% to 34.40%. The models can exploit stored image-text associations, whereas free-form bridge-name generation remains difficult.

Structural matching. Category recognition improves steadily with scale, with T2 increasing from 36.80% at 0.8B to 66.30% at 27B. T4 increases from 56.40% at 0.8B to 84.10% at 9B, while T5 increases from 59.80% at 0.8B to 75.90% at 27B and reaches 79.10% for both GPT models. Under the same reasoning budget, 35B-A3B outperforms the dense 27B model on T1 and T3 but underperforms it on T4 and T5, reflecting broader knowledge coverage alongside weaker visual reasoning in the MoE model.

Physical metric. T7 accuracy increases from 47.40% at 0.8B to 59.30% at 35B-A3B and reaches 61.20% with GPT-5.6-sol. T6 MAE increases from 190.65 m at 0.8B to 346.72 m at 4B, then decreases to 174.86 m at 27B, 189.88 m at 35B-A3B, and 75.72 m with GPT-5.6-sol. Relative height comparison improves with scale, whereas absolute height estimation follows a different trajectory.

Retrieval evidence. The lower block of Table 2 varies the retrieval configuration on a fixed Qwen3.5-4B backbone. MoRE raises T1 from 0.82% to 41.34% and T2 from 55.70% to 70.20%, while UniRAG and EchoSight reach the highest T3 full-group accuracies at 52.80% and 52.60%. Gains on visual matching are smaller and depend on the retrieval format. EchoSight reaches the highest T4

Table 2: Unified results along the progressive diagnostic chain. Upper block, ten models under closed-book conditions; lower block, five retrieval configurations on the Qwen3.5-4B backbone, with the Qwen3.5-4B row as baseline. Bold marks the best result per task within each block.

Model	T1	T2	T3 item	T3 grp	T4	T5	T6 MAE/MRE	T7
Random	0.00	12.50	26.39	7.55	25.00	50.00	—	50.00
Qwen3.5-0.8B	0.12	36.80	25.65	3.60	56.40	59.80	190.65 / 128.63	47.40
Qwen3.5-2B	0.39	53.90	30.53	9.00	70.00	73.40	277.25 / 217.00	51.60
Qwen3.5-4B	0.82	55.70	45.05	19.00	77.70	70.90	346.72 / 273.99	55.40
Qwen3.5-9B	0.82	60.20	47.32	23.30	84.10	73.10	206.94 / 152.34	58.60
Qwen3.5-27B	0.91	**66.30**	55.75	32.30	83.50	75.90	174.86 / 122.02	60.30
Qwen3.5-35B-A3B	0.97	63.70	**57.03**	**34.40**	81.80	70.60	189.88 / 136.31	59.30
Gemma-4-E2B	0.33	46.60	17.16	1.30	44.40	65.30	359.41 / 287.08	51.20
Gemma-4-E4B	0.25	48.80	22.94	4.30	64.90	66.10	127.52 / 95.13	56.90
GPT-5.5	3.67	63.60	50.92	19.90	87.10	**79.10**	111.82 / 78.00	61.00
GPT-5.6-sol	**4.59**	63.40	50.44	13.70	**88.40**	**79.10**	**75.72 / 54.53**	**61.20**
EchoSight	32.87	63.00	71.79	52.60	**80.60**	66.10	98.65 / 64.18	58.70
VisRAG-PNG	5.56	63.40	43.57	20.10	75.50	69.00	71.58 / 51.63	50.10
VisRAG-PDF	1.26	56.80	46.45	23.10	77.70	**73.70**	113.24 / 73.06	57.80
UniRAG	23.91	59.50	**74.57**	**52.80**	67.90	63.10	58.95 / 45.26	**67.70**
MoRE	**41.34**	**70.20**	70.57	48.40	76.80	70.60	**57.32 / 42.62**	65.50

Note. Qwen3.5 models are reported in (Qwen Team, 2026), Gemma-4 models in (Gemma Team et al., 2026), GPT-5.5 in (OpenAI, 2026a), GPT-5.6-sol in (OpenAI, 2026b), EchoSight in (Yan & Xie, 2024), VisRAG in (Shi et al., 2024), UniRAG in (Sharifymoghaddam et al., 2025), and MoRE in (Peng et al., 2025).

accuracy at 80.60%, VisRAG-PDF reaches the highest T5 accuracy at 73.70%, and UniRAG lowers T4 and T5 to 67.90% and 63.10%. MoRE obtains the lowest T6 MAE and UniRAG the highest T7 accuracy. Retrieval therefore improves different task dimensions through different evidence formats rather than producing a uniform gain across the benchmark.

4.3 Visual-Detail Ablation

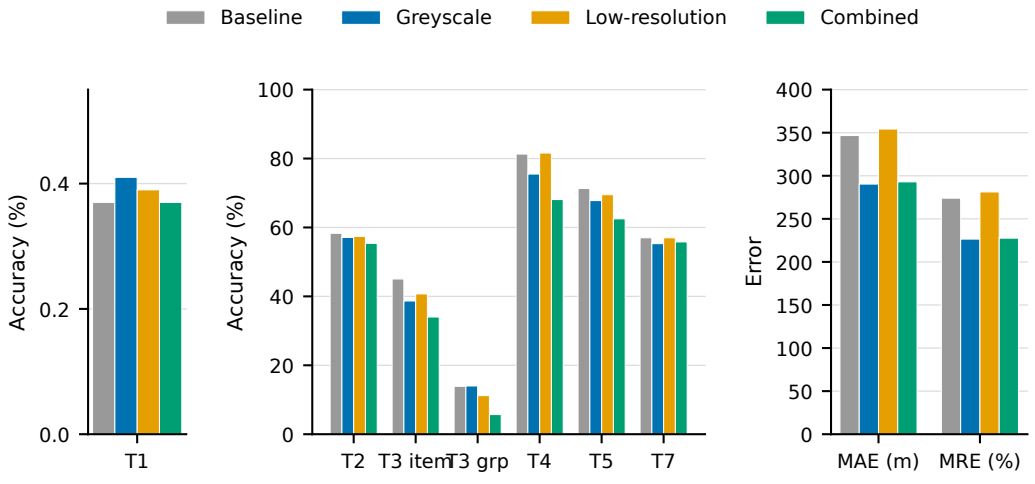

Figure 4: Visual-detail ablation for Qwen3.5-4B in bar form. Each group of four bars compares the baseline, greyscale, low-resolution, and combined conditions on one metric. Left panel, T1 name recall. Middle panel, accuracy tasks. Right panel, T6 errors, for which lower is better.

To isolate the effects of color and spatial resolution, we evaluate Qwen3.5-4B under grayscale, low-resolution, and combined transformations (Figure 4). Grayscale converts each image to a single channel and replicates it across three channels, preserving the spatial layout. Low-resolution reduces both image dimensions to one quarter and then restores the original input size by nearest-neighbor interpolation. The combined condition applies both transformations.

Relative to the full-color baseline, grayscale reduces T2 by 1.20 percentage points, T4 by 5.80 points, and T5 by 3.50 points, while T3 item-level accuracy decreases by 6.39 points and full-group accuracy remains nearly unchanged. Low-resolution reduces T2 by 0.90 points, T3 item-level accuracy by 4.34 points, T3 full-group accuracy by 2.70 points, and T5 by 1.80 points, while T4 changes by only 0.30 points. The combined condition produces the largest changes in structural matching, reducing T2 by 2.90 points, T3 full-group accuracy by 8.20 points, T4 by 13.20 points, and T5 by 8.80 points. The two transformations therefore affect the task interfaces differently rather than producing a uniform accuracy shift.

4.4 Structural Decomposition of Scale and Deviations

This subsection quantifies the geometric blind spot through G, the discrimination on same-country negatives of T5, where only structural geometry separates the two bridges. A structural equation over the task measurements yields a dominant scale factor L_M that follows a scaling law with model size M, with residual factors too small to separate from noise at ten model points, so the blind spot hides inside L_M. We decompose L_M into general capability U, image-text capability K, and geometric capability G with $L_M = U + K + G$, where G carries the blind spot. Here U is base knowledge and reasoning, K is the cross-modal alignment of images with textual knowledge, and G is the geometric cognition and comparison of geometry across image samples. Two windows measure G, both from response-level records rather than the aggregate tables; small models fall below the predicted level precisely on same-type comparisons, and four of the five retrieval configurations score below the closed book on same-type verification.

All task scores are corrected for chance and mapped through the inverse standard normal distribution, placing every task on one unbounded evidence scale, with Φ the standard normal distribution function and $z = \Phi^{-1}$ its inverse. On T5 the quantity is the signal-detection discrimination

$$d = z(r) - z(f), \tag{2}$$

with r and f the true and false positive rates on same-bridge questions, measuring how far apart the model's evidence lies for the two question classes in units of the noise, because accuracy also depends on where the model places its answer threshold, as Qwen3.5-27B and 35B-A3B show with nearly equal discrimination, 1.502 and 1.509, but a 5.3-point difference in T5 accuracy. The quantity d serves as the common currency below.

The structural equation is fitted by letting each measurement of T2, T3, T4, T5, and T7, together with the same-country and cross-country subtask scores, be predicted from the remaining measurements across the ten models, and the mutual predictions reduce to a single scale factor L_M with one loading per task; T1 and T6 are excluded and serve as external checks. Task design supplies two further quantities. On same-country negative pairs, which share architectural style and environment, the discrimination G reads the geometric ability itself, and on cross-country pairs it runs higher by the cue contribution C, hence equals $G+C$. The cross-type options of T4 are eliminated by category recognition, leaving a two-alternative choice, so a model with discrimination d should answer T4 at the rate

$$P_{\mathrm{T4}} = \Phi(d/\sqrt{2}), \tag{3}$$

and the shortfall is the T4 gap, the per-model size of the blind spot.

We next ask how G grows with M. Fitting the saturating curve

$$G(M) = A - BM^{-\kappa} \tag{4}$$

which captures fast growth at small scale slowing at large scale (Kaplan et al., 2020; Hoffmann et al., 2022), to the G series gives $A_G = 1.236$ and to the $G + C$ series $A_{G+C} = 1.997$, with residual sums of squares 0.030 and 0.233. The asymptote gap of 0.761, worth 10.9 accuracy points, does not close with scale, and A_G corresponds to a same-country accuracy ceiling of $\Phi(A_G/2) = 73.2\%$.

On the T4 prediction (Table 3), the six strong models deviate by at most 3.90 percentage points, while the four weak models fall 8.5 to 28.3 below. Qwen3.5-4B is the most special, passing with a gap of 3.2 although its overall discrimination falls below that of 2B, which fails by 13.6, so the geometric ability appears at a lower overall level than the scale series alone would suggest. Both the retrieval and degradation experiments run on this boundary model, the one asking whether external evidence can lift the newly acquired ability, the other which input signal it reads. On the country split, C rises with scale from 0.357 to 0.968, so scale enlarges the cue rather than diluting it.

The degradation conditions, in the lower block of Table 3, separate geometry from cue. Low resolution lowers G and leaves C almost unchanged; greyscale halves C and leaves G intact, so reducing resolution does lower the geometric capability while removing color does not.

Retrieval supplies the second window, measured in the middle block of Table 3 with the Qwen3.5-4B row as baseline; in all five configurations the retriever only selects content and places it in the model's context, where one attention mechanism reads it together with the query images. EchoSight and UniRAG supply bridge names, lifting G to 1.518 and 1.455, yet their T4 gaps widen from -3.2 to -7.1 and -19.2, so retrieval breaks the format transfer the boundary model had just acquired. VisRAG-PDF lifts G to 1.490 with zero T4 gain, and its PNG rendering loses further detail, with C falling to 0.252. MoRE lifts T1 most at 2.29 while its G falls to 0.894 and its C rises to 0.767, displacing geometric evidence and adding cue-relevant evidence at once. Every configuration that lifts G widens its T4 gap, so the retrieved content serves the K and cue parts and leaves L_M unchanged, and retrieval supplements evidence and cues but never geometric perception.

Table 3: Discrimination measures on T5 and derived quantities. Upper block, ten models under closed-book conditions; middle block, five retrieval configurations on the Qwen3.5-4B backbone; lower block, degradation conditions on the same backbone.

Model	d	G	C	$G{+}C$	**T4 pred**	**T4 gap**	δ_s	δ_x	Δd(T1)	**T4 same**	**T4 cross**
Qwen3.5-0.8B	0.541	0.409	0.357	0.766	64.9	-8.5	—	—	—	53.6	60.4
Qwen3.5-2B	1.383	1.172	0.661	1.834	83.6	-13.6	—	—	—	65.3	76.9
Qwen3.5-4B	1.238	1.057	0.552	1.609	80.9	-3.2	—	—	—	74.9	81.8
Qwen3.5-9B	1.382	1.237	0.443	1.680	83.6	$+0.5$	—	—	—	80.6	89.2
Qwen3.5-27B	1.502	1.281	0.700	1.981	85.6	-2.1	—	—	—	80.3	88.2
Qwen3.5-35B-A3B	1.509	1.276	0.968	2.244	85.7	-3.9	—	—	—	78.6	86.5
Gemma-4-E2B	0.852	0.813	0.087	0.899	72.7	-28.3	—	—	—	43.2	46.2
Gemma-4-E4B	1.126	1.048	0.197	1.246	78.7	-13.8	—	—	—	59.9	72.2
GPT-5.5	1.620	1.394	0.684	2.078	87.4	-0.3	—	—	—	84.0	91.6
GPT-5.6-sol	1.621	1.396	0.643	2.039	87.4	$+1.0$	—	—	—	85.7	92.4
EchoSight	1.643	1.518	0.398	1.916	87.7	-7.1	1.09	1.04	$+1.96$	76.6	86.5
UniRAG	1.599	1.455	0.523	1.978	87.1	-19.2	1.00	1.15	$+1.69$	62.6	75.7
VisRAG-PDF	1.687	1.490	0.744	2.234	88.4	-10.7	1.05	1.55	$+0.16$	73.7	83.5
VisRAG-PNG	1.356	1.268	0.252	1.520	83.1	-7.6	0.70	0.00	$+0.81$	70.3	83.0
MoRE	1.135	0.894	0.767	1.661	78.9	-2.1	-0.56	$+0.41$	$+2.18$	70.7	85.7
Full color	1.238	1.057	0.552	1.609	80.9	$+0.4$	—	—	—	74.9	81.8
Greyscale	1.160	1.062	0.273	1.335	79.4	-3.9	$+0.10$	-0.90	-0.03	70.2	74.4
Low resolution	1.019	0.876	0.408	1.284	76.4	$+5.2$	-0.59	-0.97	-0.12	73.5	84.5
Combined	0.812	0.711	0.279	0.990	71.7	-3.6	-0.78	-1.27	-0.15	60.7	70.0

5 Conclusion

BridgeGeom separates recognising a bridge from discriminating it and finds a persistent gap. A structural equation over the task measurements shows that the geometric blind spot lives inside the scaling law, and decomposing the scale factor into general, image-text, and geometric capability isolates it as a measurable quantity. This quantity saturates with model size, is never lifted by retrieval augmentation, and reads spatial detail rather than color. The blind spot is therefore a perception deficit that training objectives must address rather than a knowledge deficit that retrieval can fill, and BridgeGeom provides the controlled setting to measure whether future training closes it.

REFERENCES

Stan Antol et al. VQA: Visual question answering. In *ICCV*, 2015. URL https://doi.org/10.1109/ICCV.2015.279.

David Anugraha et al. M4-RAG: A massive-scale multilingual multi-cultural multimodal RAG. In *CVPR*, 2026. URL https://doi.org/10.48550/arXiv.2512.05959.

Zhen Cai et al. DepthLM: Metric depth from vision language models. In *ICLR*, 2026. URL https://doi.org/10.48550/arXiv.2509.25413.

R. F. Çelik et al. Vision language models for bridge inspections: A review of applications in image-based damage documentation. *Structural Concrete*, 2026. URL https://doi.org/10.1002/suco.70702.

Zhe Cheng et al. V-STaR: Benchmarking video-LLMs on spatio-temporal reasoning. *arXiv preprint arXiv:2503.11495*, 2025. URL https://doi.org/10.48550/arXiv.2503.11495.

Oskar Delzer and Silke Bender. Reproducibility study on how to find spurious correlations, shortcut learning, clever hans or group-distributional non-robustness and how to fix them. *arXiv preprint arXiv:2604.04518*, 2026. URL https://doi.org/10.48550/arXiv.2604.04518.

Mark Endo and Stephanie Yeung-Levy. Downscaling intelligence: Exploring perception and reasoning bottlenecks in small multimodal models. In *CVPR*, 2026. URL https://doi.org/10.48550/arXiv.2511.17487.

Chaoyou Fu et al. MME: A comprehensive evaluation benchmark for multimodal large language models. In *NeurIPS Datasets and Benchmarks*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023.

Kartik Garg et al. Revisit anything: Visual place recognition via image segment retrieval. In *ECCV*, 2024. URL https://doi.org/10.1007/978-3-031-73113-6_19.

Robert Geirhos, Jonas-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *arXiv preprint arXiv:2004.07780*, 2020. URL https://doi.org/10.48550/arXiv.2004.07780.

Gemma Team et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026. URL https://doi.org/10.48550/arXiv.2607.02770.

Shuo Gong et al. Learning hierarchical and geometry-aware graph representations for Text-to-CAD. In *ICLR*, 2026. URL https://doi.org/10.48550/arXiv.2604.10075.

Weizhen He et al. Instruct-ReID: A multi-purpose person re-identification task with instructions. In *CVPR*, 2024. URL https://doi.org/10.1109/CVPR52733.2024.01659.

HighestBridges. Highestbridges database. https://highestbridges.com, 2025.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022. URL https://doi.org/10.48550/arXiv.2203.15556.

Wen Hu et al. Beyond geo-localization: Fine-grained orientation of street-view images by cross-view matching with satellite imagery. In *ACM Multimedia*, 2022. URL https://doi.org/10.1145/3503161.3548102.

Drew Hudson and Christopher Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. URL https://doi.org/10.1109/CVPR.2019.00686.

Yao Jia et al. Benchmarking multimodal knowledge conflict for large multimodal models. In *AAAI*, 2026. URL https://doi.org/10.48550/arXiv.2505.19509.

Justin Johnson et al. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017. URL https://doi.org/10.1109/CVPR.2017.215.

Jared Kaplan, Sam McCandlish, Tom Henighan, et al. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. URL https://doi.org/10.48550/arXiv.2001.08361.

Jonathan Krause et al. 3D object representations for fine-grained categorization. In *ICCV Workshops*, 2013. URL https://doi.org/10.1109/ICCVW.2013.77.

Ola Kristoffersen et al. MMLandmarks: A cross-view instance-level benchmark for geo-spatial understanding. In *CVPR*, 2026. URL https://doi.org/10.48550/arXiv.2512.17492.

Bo Li et al. SEED-Bench: Benchmarking multimodal LLMs with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a. URL https://doi.org/10.48550/arXiv.2307.16125.

Lei Li et al. The last visible pixel: Probing fine-scale perception in vision-language models. *arXiv preprint arXiv:2606.07861*, 2026. URL https://doi.org/10.48550/arXiv.2606.07861.

Sheng Li et al. CLIP-ReID: Exploiting vision-language models for image re-identification without concrete text labels. In *AAAI*, 2023b. URL https://doi.org/10.1609/aaai.v37i1.25225.

Fang Lin et al. Do vision-language models measure up? benchmarking visual measurement reading with MeasureBench. In *CVPR*, 2026. URL https://doi.org/10.48550/arXiv.2510.26865.

Yuan Liu et al. MMBench: Is your multi-modal model an all-around player? In *ECCV*, 2024. URL https://doi.org/10.1007/978-3-031-72658-3_13.

Michael A. Lones. How to avoid machine learning pitfalls: A guide for academic researchers. *arXiv preprint arXiv:2108.02497*, 2021. URL https://doi.org/10.48550/arXiv.2108.02497.

Tao Lu et al. OmniMech: All-in-one multimodal mechanical benchmark for 3D reconstruction. *arXiv preprint arXiv:2608.05539*, 2026. URL https://doi.org/10.48550/arXiv.2608.05539.

Xingyang Lu et al. GLEAM: Learning to match and explain in cross-view geo-localization. *arXiv preprint arXiv:2509.07450*, 2025a. URL https://doi.org/10.48550/arXiv.2509.07450.

Yiming Lu et al. LLaVA-ReID: Selective multi-image questioner for interactive person re-identification. *arXiv preprint arXiv:2504.10174*, 2025b. URL https://doi.org/10.48550/arXiv.2504.10174.

Shuo Luo et al. GeoGramBench: Benchmarking the geometric program reasoning in modern LLMs. In *ICLR*, 2026. URL https://doi.org/10.48550/arXiv.2505.17653.

Xin Ma et al. Do 3D large language models really understand 3D spatial relationships? In *ICLR*, 2026. URL https://doi.org/10.48550/arXiv.2603.23523.

Subhransu Maji et al. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. URL https://doi.org/10.48550/arXiv.1306.5151.

Damiano Marsili et al. Same or not? enhancing visual perception in vision-language models. In *CVPR*, 2026. URL https://doi.org/10.48550/arXiv.2512.23592.

OpenAI. GPT-5.5 system card. Technical report, OpenAI, April 2026a. URL https://openai.com/index/gpt-5-5-system-card/.

OpenAI. GPT-5.6 system card. Technical report, OpenAI, July 2026b. URL https://openai.com/index/gpt-5-6/.

Maxime Oquab et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. URL https://doi.org/10.48550/arXiv.2508.10104.

Cheng Peng et al. Mixture-of-retrieval experts for reasoning-guided multimodal knowledge exploitation. *arXiv preprint arXiv:2505.22095*, 2025. URL https://doi.org/10.48550/arXiv.2505.22095.

Jing Qiu et al. Semantic richness or geometric reasoning? the fragility of VLM's visual invariance. *arXiv preprint arXiv:2604.01848*, 2026. URL https://doi.org/10.48550/arXiv.2604.01848.

Qwen Team. Qwen3.5: Towards native multimodal agents. Qwen Blog, https://qwen.ai/blog?id=qwen3.5, 2026.

Sahel Sharifymoghaddam et al. UniRAG: Universal retrieval augmentation for large vision-language models. In *Findings of NAACL*, 2025. URL https://doi.org/10.18653/v1/2025.findings-naacl.108.

Yu Shi et al. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*, 2024. URL https://doi.org/10.48550/arXiv.2410.10594.

Yevgeny P. Shkolnikov. Do foundation models know geometry? probing frozen features for continuous physical measurement. *arXiv preprint arXiv:2603.06459*, 2026. URL https://doi.org/10.48550/arXiv.2603.06459.

Manav Talreja et al. GeoRC: A benchmark for geolocation reasoning chains. In *ACL*, 2026. URL https://doi.org/10.18653/v1/2026.acl-long.1883.

Tristan Thrush et al. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*, 2022. URL https://doi.org/10.1109/CVPR52688.2022.00517.

Théo Vitry et al. Bias leaves a gradient trail: Label-free bias identification via gradient probes on concept decompositions. In *KI*, 2026. URL https://doi.org/10.48550/arXiv.2605.28780.

Xing Wang et al. Fine-grained cross-view geo-localization using a correlation-aware homography estimator. In *NeurIPS*, 2023. URL https://doi.org/10.52202/075280-0234.

Ahsan T. Wasi et al. SpatiaLab: Can vision-language models perform spatial reasoning in the wild? In *ICLR*, 2026. URL https://doi.org/10.48550/arXiv.2602.03916.

Tao Weng et al. OddGridBench: Exposing the lack of fine-grained visual discrepancy sensitivity in multimodal large language models. In *CVPR*, 2026. URL https://doi.org/10.48550/arXiv.2603.09326.

Hao Xu et al. Beyond visual similarity: Entity-aligned retrieval for knowledge-based visual question answering. In *ACM Multimedia*, 2026a. URL https://doi.org/10.48550/arXiv.2608.21450.

Wenqi Xu et al. VisuLogic: A benchmark for evaluating visual reasoning in multi-modal large language models. *arXiv preprint arXiv:2504.15279*, 2025. URL https://doi.org/10.48550/arXiv.2504.15279.

Yang Xu et al. Diagnosing dense same-class attribute misbinding in large vision-language models. *arXiv preprint arXiv:2608.16805*, 2026b. URL https://doi.org/10.48550/arXiv.2608.16805.

Yifan Yan and Wenxuan Xie. EchoSight: Advancing visual-language models with wiki knowledge. In *Findings of EMNLP*, 2024. URL https://doi.org/10.18653/v1/2024.findings-emnlp.83.

Chen Yang et al. Thinking in structures: Evaluating spatial intelligence in constraint-governed spaces. In *ICML*, 2026. URL https://doi.org/10.48550/arXiv.2602.07864.

Sen Yang et al. MMSI-Bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*, 2025. URL https://doi.org/10.48550/arXiv.2505.23764.

Songtao Yu et al. How far are VLMs from visual spatial intelligence? *arXiv preprint arXiv:2509.18905*, 2025. URL https://doi.org/10.48550/arXiv.2509.18905.

Xiang Yue et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *CVPR*, 2024. URL https://doi.org/10.1109/CVPR52733.2024.00913.

Chen Zhang et al. Towards mixed-modal retrieval for universal retrieval-augmented generation. In *ACM Web Conference*, 2026a. URL https://doi.org/10.48550/arXiv.2510.17354.

Qiang Zhang et al. Probing and bridging geometry-interaction cues for affordance reasoning in vision foundation models. In *CVPR*, 2026b. URL https://doi.org/10.48550/arXiv.2602.20501.

Zhi Zhang et al. Benchmarking and enhancing VLM for compressed image understanding. In *ICML*, 2026c. URL https://doi.org/10.48550/arXiv.2512.20901.

Yang Zheng et al. Dual-pathway geometry-aware MLLM for spatial intelligence. *arXiv preprint arXiv:2605.25334*, 2026. URL https://doi.org/10.48550/arXiv.2605.25334.

Hao Zhong et al. Can large multimodal models inspect buildings? a hierarchical benchmark for structural pathology reasoning. *arXiv preprint arXiv:2603.20148*, 2026. URL https://doi.org/10.48550/arXiv.2603.20148.

Jialong Zuo et al. UFineBench: Towards text-based person retrieval with ultra-fine granularity. In *CVPR*, 2024. URL https://doi.org/10.1109/CVPR52733.2024.02078.

AI USAGE DISCLOSURE

This work used AI tools in the following three areas, with all scientific decisions made by the authors.

Writing assistance. The GLM 5.3 model was used to draft and polish the manuscript, including compressing text to meet page limits and maintaining consistency of notation across sections. All experimental design, data analysis, and interpretation of results were conducted solely by the authors.

Retrieval and discovery. The GLM 5.3 model was used with web search to discover and verify citations for related work, particularly for publications from 2026. Each citation was manually confirmed against its original source before inclusion. The survey scope, field narrative, and the decision of which work to cite or exclude were determined by the authors.

Data generation and annotation. GPT-5.4 was used in the dataset construction pipeline in two roles described in Section 3.1, first as an automated filter that labels whether each image is a real bridge photograph, and second as a category annotator that assigns each bridge to one of eight reference categories over five independent runs with majority vote. DINOv3 was used to encode image features for the fine-filtering stage. An independent language model was used as judge for the T1 alias evaluation reported in Appendix B. No AI tool was used to generate bridge images, and no evaluated model participated in the construction of the benchmark it was later scored on.

A DATASET STATISTICS

Table 4: Dataset composition by reference category.

Category	Bridges	Images
Beam	429	3,299
Arch	200	1,485
Footbridge	144	867
Cable-stayed	93	774
Suspension	73	636
Truss	18	112
Cantilever	10	73
Pipeline	4	26
Total	971	7,555

Country metadata covers 971 bridges across 32 countries, of which 736 (75.80%) are in China, 66 in Japan, 61 in Italy, 17 in the United States, and 11 in Mexico. Deck-to-ground heights range from 30 to 625 metres with a median of 130. Each bridge has between 5 and 27 retained images, with a median of 6 and a mean of 7.78.

B T1 Dual Evaluation

T1 exact-match scoring penalizes aliasing, transliteration, and partial names. We therefore evaluate T1 under a second criterion using an independent language model as judge. The judge receives the gold bridge name and the model's raw output, and rules whether the two refer to the same bridge. Translations, transliterations, spelling variants, and common aliases count as the same bridge; site replacements, rail-road variants, and same-type but different bridges count as different. Table 5 reports both criteria for all models and retrieval configurations.

Table 5: T1 bridge-name recall under exact match and judge-verified criteria.

Model / Config	Exact	Judge	Gain
Qwen3.5-0.8B	0.12	1.50	+1.38
Qwen3.5-2B	0.39	3.91	+3.52
Qwen3.5-4B	0.82	5.09	+4.26
Qwen3.5-9B	0.82	4.68	+3.85
Qwen3.5-27B	0.91	3.58	+2.68
Qwen3.5-35B-A3B	0.97	5.05	+4.08
Gemma-4-E2B	0.33	1.98	+1.65
Gemma-4-E4B	0.25	1.17	+0.93
GPT-5.5	3.67	12.07	+8.40
GPT-5.6-sol	4.59	13.80	+9.21
Grok 4.5†	0.62	1.65	+1.03
EchoSight	32.87	36.99	+4.12
VisRAG-PNG	5.56	7.97	+2.41
VisRAG-PDF	1.26	5.66	+4.41
UniRAG	23.91	26.92	+3.01
MoRE	41.34	43.32	+1.98

Note. Values in percent. †Grok 4.5 is evaluated on a 10% canonical subset (486 items); all others use the full 4,855 items.

The judge-verified scores show that models do store bridge identities but cannot reliably produce the exact name string. GPT-5.6-sol reaches 13.80% under the judge criterion versus 4.59% under exact match, and MoRE reaches 43.32% versus 41.34%. The gap is driven by aliases and translations, such as Wuzuohe Bridge answered as Wuzuo River Bridge, Fades Railway Viaduct answered as Viaduc des Fades, and Ponte della Vittoria answered in Italian. Under both criteria the ranking is unchanged, so the exact-match scores reported in the main text underestimate name recall without affecting any conclusion.

C Supplementary Results

Grok 4.5. Grok 4.5 is evaluated on a 10% canonical subset (1,083 items) and reaches 63.00% on T2, 58.08% on T3 item accuracy, 57.00% on T4, and 65.00% on T5, consistent with the mid-range open-source models.

MRAG gains. Table 6 reports the per-task gain of each retrieval configuration over the closed-book baseline on Qwen3.5-4B.

Table 6: MRAG gains over closed-book Qwen3.5-4B, in percentage points.

	Echo	PNG	PDF	UniRAG	MoRE
T1 (exact)	+32.05	+4.74	+0.43	+23.09	+40.51
T2	+7.30	+7.70	+1.10	+3.80	+14.50
T3 item	+26.74	−1.48	+1.40	+29.52	+25.52
T3 group	+33.60	+1.10	+4.10	+33.80	+29.40
T4	+2.90	−2.20	+0.00	−9.80	−0.90
T5	−4.80	−1.90	+2.80	−7.80	−0.30
T7	+3.30	−5.30	+2.40	+12.30	+10.10

The delta table makes the asymmetry visible at a glance. Every configuration that gains on T1 and T3 simultaneously fails or barely moves on T4 and T5, with UniRAG losing 9.80 points on T4 and 7.80 points on T5. This is the accuracy-level counterpart of the discrimination-level finding in Section 4.4.

D DEGRADATION FULL TABLE

Table 7: Degradation ablation on Qwen3.5-4B. Accuracy in percent, T6 MAE in metres.

Condition	**T1**	**T2**	**T3 item**	**T3 grp**	**T4**	**T5**	**T6**	**T7**
Full color	0.82	55.70	45.05	19.00	81.30	70.90	346.72	55.40
Greyscale	0.76	54.50	38.66	19.20	75.50	67.40	346.10	54.50
Low resolution	0.58	54.80	40.71	16.30	81.60	69.10	394.17	54.00
Combined	0.54	52.80	34.05	10.80	68.10	62.10	469.10	52.10