Explaining Human Preferences via Metrics for Structured 3D Reconstruction

Jack Langerman
Independent Researcher [*]
jack@jackml.com

Denys Rozumnyi
ETH Zurich / Faculty of Electrical Engineering, CTU in Prague [†]
rozumden@gmail.com

Yuzhong Huang
HOVER Inc.
yuzhong.huang@hover.to

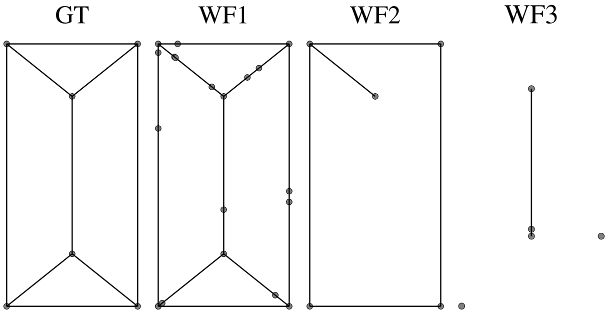
Dmytro Mishkin
HOVER Inc. / Faculty of Electrical Engineering, CTU in Prague
dmytro.mishkin@hover.to

Abstract

”What cannot be measured cannot be improved” while likely never uttered by Lord Kelvin, summarizes effectively the driving force behind this work. This paper presents a detailed discussion of automated metrics for evaluating structured 3D reconstructions. Pitfalls of each metric are discussed, and an analysis through the lens of expert 3D modelers’ preferences is presented. A set of systematic ”unit tests” are proposed to empirically verify desirable properties, and context aware recommendations regarding which metric to use depending on application are provided. Finally, a learned metric distilled from human expert judgments is proposed and analyzed. The source code is available at https://github.com/s23dr/wireframe-metrics-iccv2025.

1. Introduction

Benchmarks have been key drivers of progress in computer vision; the canonical example is certainly ImageNet [31], but prominent examples abound beyond image classification: object tracking [21, 23], image retrieval [4, 30], image matching [20], 6D pose estimation [15], optical flow [27], *etc*. Benchmarks have three main components – the data, the protocol, and the metrics. While the data is the single most important component, progress is hard without being able to answer the question, ”progress on what?” Good metrics are the quantitative answer to this question. While metrics do not need to be perfect, their gradient should point progress in the right direction.



Metric/wireframe	GT	WF1	WF2	WF3
Vertex F1 ↑	1.00	0.56	0.91	0.18
Edge F1 ↑	1.00	0.19	0.71	0.00
Jaccard Distance ↓	0.00	0.00	0.33	1.00
WED ↓	0.00	2.32	0.52	1.82
WED S23DR ↓	0.00	2.69	0.59	1.63
Graph Spectral ↓	0.00	377.93	577.49	1603.32

Figure 1. A motivating example for this work. While humans tend to sort the wireframes from best to worst in the presented order, popular metrics (defined in Sec 3) sort them differently, sometimes completely inverting the order. Top, left to right: GT – ground truth wireframe, WF1 – wireframe with edges split into several segments, maintaining geometrical and topological accuracy, WF2 – wireframe with missing vertices and edges, WF3 – wireframe with only one correct vertex. Bottom: distances between GT and respective wireframe. Numbers that change sorting are in red.

We consider an area of structured and semi-structured reconstruction, which has recently gained popularity [1] [19] [5]. Given a set or sequence of sensory data, such as ground images [22], satellite images [9], or aerial LiDAR [33], the goal is to produce a wireframe or a CAD model of a building or other structure. The modeling output is presented as a spatial graph with vertices (such as roof apex point, *etc*.) and edges (ridge line, *etc*.). Several datasets have been recently proposed [22, 33] for the task. The issue is that hardly a pair of publications in the area use the same metric to evaluate quantitative results. Some use recognition metrics such as precision and recall on vertices and edges [25] or more enhanced versions such as Structured Average Precision [26, 36]. Others opt for graph-based metrics, such as the Wireframe Edit Distance (WED) [8, 24, 25]. Finally, some methods treat the problem similarly to point cloud registration and report Chamfer

[*] Now at Apple
[†] Now at Meta

Distance (CD) [10, 16, 17]. Other related fields, like structure from motion, use downstream metrics, such as image generation quality [6] and camera pose accuracy [20]. One aspect of the difficulty is related to the fact that structured reconstructions often have different goals. For instance, one purpose of the reconstructed wireframes is to represent building plans and answer questions such as ”what is the area of the bedroom?” Within this formulation, a black-box model, *e.g.* visual-and-language model (VLM), which takes an image as an input and outputs a correct estimate, would be a perfect match. On the other hand, a floorplan or a blueprint has value for record-keeping, planning, and other applications that cannot be easily replaced with a black-box model.

Finally, many existing metrics, while useful, often fail to deliver value in practice when comparing two imperfect estimations, and designing metrics that effectively compare a pair of ”very good” and a pair of ”very bad” solutions at the same time is also non-trivial.

Moreover, there is evidence [22] that some metrics can be ”hacked” or exploited in such a way that obviously bad solutions have better scores than flawed but ultimately quite reasonable solutions. Such examples include a number of corner cases, which can cause existing metrics to become useless in practical scenarios. For instance, in the case where long edges are split into smaller, yet perfectly collinear, pieces, *e.g.* Fig. 1, most commonly used metrics, such as edge F1, vertex F1, and Wireframe Edit Distance, fail completely and prefer much worse solutions.

This paper makes the following contributions:

(1) Measure the perceived quality of reconstructions by human domain experts, and infer a global ranking of all structured reconstructions. This ranking is then compared to the ranking given by all metrics.

(2) Show how well existing metrics agree with human preferences and how much they correlate with each other.

(3) Propose a set of ”unit-tests” for testing the properties of the metrics.

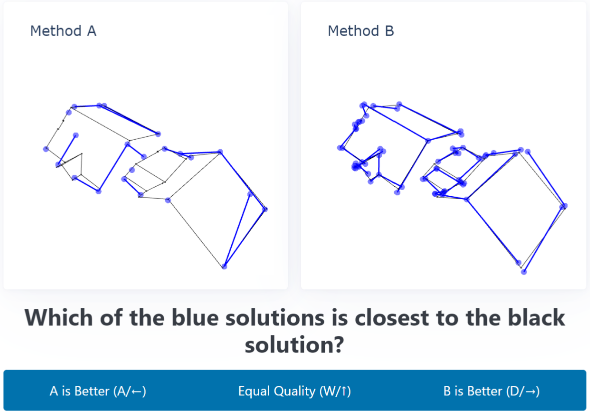
(4) Introduce a simple learned metric, which correlates well with human judgment.

(5) Make recommendations of which metrics to use depending on the use case.

2. What do we actually need from wireframe comparison metrics?

We consider semi-automated 3D modeling as a target task with enough generality to drive progress that can smoothly scale to fully-automated modeling, but is still tractable and reliable for commercial applications today.

For this task, the wireframe representation is estimated from some ”raw” inputs such as images or a LiDAR point cloud and then transferred to human experts to (1) approve it, (2) correct it and then approve, (3) reject and create the model from scratch manually.



Which of the blue solutions is closest to the black solution?

Figure 2. Wireframe ranking interface for human annotators.

We argue that such task formulation creates an implicit usefulness ranking over reconstructions (and thereby over reconstruction methods). Otherwise, without considering human involvement, everything becomes binary – either the model is good enough to be used (*e.g.* for 3D printing, measurement extraction), or it is not.

For this reason, we consider the following experimental setup to benchmark wireframe comparison metrics. A pool of professional 3D modelers, whose everyday job is creating CAD-like models from raw data such as images, is asked to rank pairs of wireframes. An example of the ranking setup is shown in Fig. 2. All wireframes are superimposed with ground truth models, and annotators are provided tools allowing them to translate, scale, and rotate the wireframes in 3D.

The wireframes are drawn from two pools, as described below. We then evaluate how the existing metrics agree with the judgment of professional human 3D modeling experts.

Pool1– $S^2 3DR$. We acquire a representative set of $S^2 3DR$ challenge entries [22] as well as a PC2WF [24] baseline. These wireframes were algorithmically (and with the help of deep-learning models) reconstructed from multiview inputs with the goal of minimizing a variant of WED. We include the top-10 entries with team names used as identifiers. The ground truth models were created by human experts and have undergone significant validation. The input data were captured by users on mobile phones in North America.

Pool2 — Corrupted ground truth. We apply one of the following operations on the ground-truth wireframes from Pool 1 – examples are shown in Fig. 3:

- (deform_{low, medium, high}) Split the edge into several edges and perturb the positions of the vertices without breaking the topology.
- (perturb_{low, medium, high}) Split the vertex into sev-

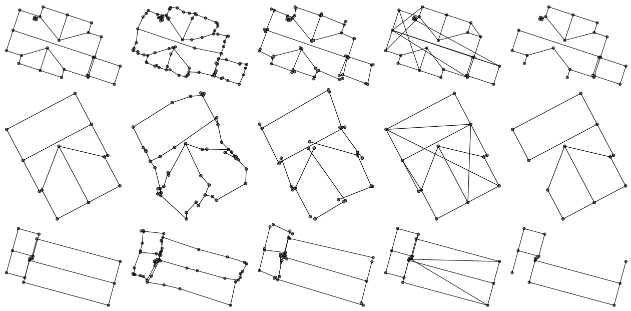


Figure 3. Examples of corrupted ground truth wireframes, used for wireframe ranking. Left to right: GT, deformed edges (deform_medium), vertex duplication and random movement (perturb_medium), edge addition (add_low), edge deletion(remove_low).

eral ones. Each of the new vertices is randomly shifted from the ground truth. If the original vertex was connected to multiple neighbors, randomly decide which of the new vertices is connected to which neighbors.

- (add_{low, medium, high}) Randomly add wrong edges to the model.
- (remove_{low, medium, high}) Randomly delete some of the vertices and all the edges connected to them.

Unit-tests & Desired properties of dissimilarity scores: In addition to being aligned with human judgment, we also propose a set of ”unit-tests” for the metrics, which we believe are reasonable, and check if the dissimilarity scores satisfy these requirements. We design tests for the formal properties of mathematical metrics as well as additional properties relevant to evaluating dissimilarity in structured reconstruction tasks: For example, if wrong edges E_1 and E_2 are added to the GT model, the metric should score the resulting wireframe lower than if E_1 or E_2 are added separately.

Identity of Indiscernibles: This property ensures that identical inputs receive a dissimilarity score of zero, indicating perfect similarity. For any reconstruction x, a metric d satisfies this property if $d(x, x) = 0$.

Symmetry: A symmetric metric produces the same dissimilarity score regardless of the order of the inputs. For reconstructions x and y, a metric satisfies symmetry if $d(x, y) = d(y, x)$.

Triangle Inequality: The triangle inequality ensures that for any three reconstructions x, y, and z, the dissimilarity between x and z is less than or equal to the sum of dissimilarities between x and y, and y and z. This relationship is expressed as $d(x, z) \leq d(x, y) + d(y, z)$.

Monotonicity: This property describes how the dissimilarity score behaves when components (such as vertices or edges) are removed from a reconstruction. A metric satisfies monotonicity if the dissimilarity score does not increase when wrong vertices or edges are deleted. Similarly, the dissimilarity must not increase when correct vertices or edges are added.

Quasi-proportionality: This property holds when the metric changes smoothly under perturbations. This is evaluated by moving random vertices with small increments and checking the variance of the differences in the score. We use the following perturbations to simulate better or worse reconstructions: (i) remove correct edges from the ground truth wireframe; (ii) add wrong edges to the ground truth wireframe; (iii) disconnect ground truth edges; (iv) remove correct vertices; (v) move ground truth vertices to the wrong location. For every perturbation, we apply it 10 times and declare an example monotonic if it is strictly increasing (or decreasing as appropriate) for those continuous 10 perturbations.

3. Metrics

The following metrics are considered:

WED – Wireframe Edit Distance was proposed by Liu *et al.* [24] as an extension of the Graph Edit Distance (GED) [32]. GED quantifies the distance between two graphs as the minimum number of elementary operations (inserting and deleting edges and vertices) required to transform one graph into another. WED extends this to wireframes (graphs with node positions and edge lengths) and proposes a cheap approximation to the NP-Hard problem of computing the optimal sequence of edits. Concretely, an assignment is first computed between the predicted and ground-truth vertices, and a cost is paid proportional to the distance between matched vertices. Next, unmatched vertices are deleted, and missing vertices are inserted (paying a cost proportional to the number of inserted/deleted vertices). Finally, given the vertex assignments, missing edges are inserted and extra edges deleted, paying a cost proportional to their length. In order to use WED, one needs to decide on the cost of insertion and deletion of the vertices, as well as the order of operations, and method of computing vertex/edge assignment. WED was used to determine the winner in the Building3D [33] and $S^2 3DR$ [22] CVPR Challenges.

ECD – Edge Chamfer Distance. ECD is commonly used in structured reconstruction papers [10, 16, 17]. We consider a family of chamfer-like metrics between two point sets A and B sampled from wireframe edges. The general form is:

$$d(A, B) := \inf_{\pi_{AB}:A\to B} \mathbb{E}_{a\in A}\left[f(a, \pi_{AB}(a))\right], \qquad (1)$$

where π_{AB} represents an assignment from elements in A to elements in B, and f is typically an ℓ_p norm of the difference between the inputs. Different constraints on π_{AB} yield different metrics:

- The classical chamfer distance corresponds to $\pi_{AB}(a) = \arg\min_{b \in B} f(a, b)$, *i.e.* nearest neighbor matching.
- The most constrained version requires π_{AB} to be a bijective matching, which can be computed via the Hungarian algorithm and is equivalent to the Earth Mover's Distance when f is the ℓ_p norm.

Length Weighted Spectral Graph Distance – SD incorporates both topological and geometric information by framing graph (wireframe) distance in terms of distances between the spectra of weighted graph Laplacians. We measure the spectral distance using the 2-Wasserstein metric between the eigenvalue distributions:

$$SD(G_1, G_2) := W_2(\lambda(L_1), \lambda(L_2)) \,, \tag{2}$$

where $\lambda(L)$ denotes the spectrum of the Laplacian L. For a graph $G = (V, E)$, the weighted graph Laplacian is defined:

$$L := D - A \tag{3}$$

where D is the weighted degree matrix ($|V| \times |V|$ diagonal matrix with each diagonal entry containing the sum of the lengths of edges incident to that vertex), and A is the weighted adjacency matrix ($|V| \times |V|$ with $A_{ij} = \|\text{coord}(V_i) - \text{coord}(V_j)\|_2$ if $(i, j) \in E$ and 0 otherwise).

Corner and Edge Metrics. We also compute precision, recall, and F1 scores for both corners and edges. For corners, we consider a prediction correct if it lies within a distance threshold of a ground truth corner. For edges, we use the Hausdorff distance between line segments to determine matches. These metrics provide an intuitive measure of the topological accuracy of the predicted wireframes.

Hausdorff Distance measures the maximum of minimal distances between two sets of points. For wireframes, we sample points along the edges and compute the Hausdorff distance between these point sets, providing a measure of geometric similarity that considers both corner positions and edge geometry.

Intersection over Union is a popular metric in a wide range of fields, *e.g.* segmentation, tracking, object detection. However, it is rarely used to assess the quality of wireframe reconstructions. We extend the definition of the wireframe as a set of cylinders with a fixed radius (the only hyperparameter of this metric) and define the metric as an IoU between two sets of cylinders, given by two wireframe reconstructions that need to be compared. An approximation via point sampling is considered: sample random points from both sets of cylinders and compute the average number of times when the point falls inside of both sets of cylinders. The **Jaccard distance** is reported between two sets.

Visual-and-language models could potentially be used for this task. We consider 4o, o1 [13], Grok 2 [34], qwen-2.5 [18], pixtral 12b [2], claude 3.5 and 3.7 [3], gemini 2.0 and gemini 2.0-flash [12] models via OpenRouter [28]. The prompts are provided in the supplementary.

Learned metric. We also explore how to distill the human annotations directly into a metric. To this end, we propose learning a metric with transfer learning. First, reconstruction and ground-truth wireframes are plotted in 3D and rendered from a canonical viewpoint (denoted r_i). Then, DiNOv2 [29] features are extracted from those renderings and an MLP-based regression head is then trained to regress scores based on the extracted features ($g(r_i)$. A Bradley–Terry [7] probability model is assumed and pairwise annotations are used to supervise the training by minimizing a binary cross-entropy loss with a batch size of 16. 10-fold cross validation splits the data such that the sets of ground truth structures and reconstruction methods used in the training and test sets are disjoint. We observe average accuracy across the folds of 76% where a prediction is considered correct if $g(r_{\text{winner}}) > g(r_{\text{loser}})$.

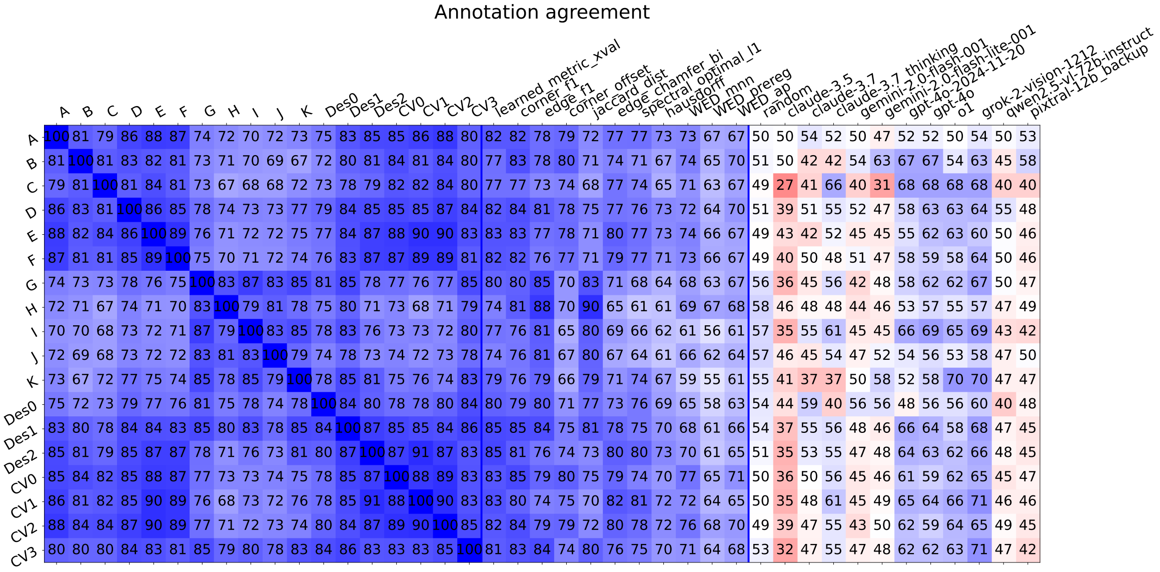
4. Experiments

Human wireframe ranking and its consistency. To determine which of the metrics under consideration are most appropriate, we employ three groups of people to provide pseudo-ground-truth rankings of the solutions. The first and biggest group is made up of human 3D modeling experts who professionally create CAD models of objects from photos. The second group is computer vision researchers, and finally, the third group is people who do not work with 3D modeling in their daily lives (designers). There are 11, 4, and 3 people in those groups respectively.

Annotators were shown pairs of reconstructions of the same structure, and asked to specify which reconstruction most closely matched the superimposed corresponding wireframe. An example of the user interface is shown in Fig. 2. Annotators were able to zoom, pan, and rotate, allowing them to examine the solutions from all sides if needed; the viewpoint of both solutions is synchronized to ease the comparison. All 27 methods are compared exhaustively by every rater, meaning $\binom{27}{2} = 351$ method pairs per house × 10 houses = 3510 pairs *for every rater*. Most raters rate a few more pairs because of the self-consistency checks. See the Suppl. for the additional information.

Rater Reliability. We quantify rater reliability using two complementary methods: self-consistency and correctness on synthetic samples.

Correctness on Synthetic Samples. We introduced synthetic pairs of wireframes with known ground-truth rankings based on systematically applied ”corruptions.” Each corruption type featured ”Low,” ”Medium,” and ”High” severity levels. We treat the ”low” vs ”high” per each corruption type to be obvious enough that if annotators rank them differently, it can be treated as a labeling error, *e.g.*



	A	B	C	D	E	F	G	H	I	J	K	Des0	Des1	Des2	CV0	CV1	CV2	CV3	learned_metric_xval	corner_f1	edge_f1	corner_offset	jaccard_dist	edge_chamfer_bi	spectral_optimal_l1	hausdorff	WED_mnn	WED_prereg	WED_ap	random	claude-3.5	claude-3.7	claude-3.7_thinking	gemini-2.0-flash-001	gemini-2.0-flash-lite-001	gpt-4o-2024-11-20	gpt-4o	o1	grok-2-vision-1212	qwen2.5-vl-72b-instruct	pixtral-12b_backup
A	100	81	79	86	88	87	74	72	70	72	73	75	83	85	85	86	88	80	82	82	78	79	72	77	77	73	73	67	67	50	50	54	52	50	47	52	52	50	54	50	53
B	81	100	81	83	82	81	73	71	70	69	67	72	80	81	84	81	84	80	77	83	78	80	71	74	71	67	74	65	70	51	50	42	42	54	63	67	67	54	63	45	58
C	79	81	100	81	84	81	73	67	68	68	72	73	78	79	82	82	84	80	77	77	73	74	68	77	74	65	71	63	67	49	27	41	66	40	31	68	68	68	68	40	40
D	86	83	81	100	86	85	78	74	73	73	77	79	84	85	85	85	87	84	82	84	81	78	75	77	76	73	72	64	70	51	39	51	55	52	47	58	63	63	64	55	48
E	88	82	84	86	100	89	76	71	72	72	75	77	84	87	88	90	90	83	83	83	77	78	71	80	77	73	74	66	67	49	43	42	52	45	45	55	62	63	60	50	46
F	87	81	81	85	89	100	75	70	71	72	74	76	83	87	87	89	89	81	82	82	76	77	71	79	77	71	73	66	67	49	40	50	48	51	47	58	59	58	64	50	46
G	74	73	73	78	76	75	100	83	87	83	85	81	85	78	77	76	77	85	80	80	85	70	83	71	68	64	68	63	67	56	36	45	56	42	48	58	62	62	67	50	47
H	72	71	67	74	71	70	83	100	79	81	78	75	80	71	73	68	71	79	74	81	88	70	90	65	61	61	69	67	68	58	46	48	48	44	46	53	57	55	57	47	49
I	70	70	68	73	72	71	87	79	100	83	85	78	83	76	73	73	72	80	77	76	81	65	80	69	66	62	61	56	61	57	35	55	61	45	45	66	69	65	69	43	42
J	72	69	68	73	72	72	83	81	83	100	79	74	78	73	74	72	73	78	74	76	81	67	80	67	64	61	66	62	64	57	46	45	54	47	52	54	56	53	58	47	50
K	73	67	72	77	75	74	85	78	85	79	100	78	85	81	75	76	74	83	79	76	79	66	79	71	74	67	59	55	61	55	41	37	37	50	58	52	58	70	70	47	47
Des0	75	72	73	79	77	76	81	75	78	74	78	100	84	80	78	78	80	84	80	79	80	71	77	73	76	69	65	58	63	54	44	59	40	56	56	48	56	56	60	40	48
Des1	83	80	78	84	84	83	85	80	83	78	85	84	100	87	85	85	84	86	85	85	84	75	81	78	75	70	68	61	66	54	37	55	56	48	46	66	64	58	68	47	45
Des2	85	81	79	85	87	87	78	71	76	73	81	80	87	100	87	91	87	83	85	81	76	74	73	80	80	73	70	61	65	51	35	53	55	47	48	64	63	62	66	48	45
CV0	85	84	82	85	88	87	77	73	73	74	75	78	85	87	100	88	89	83	83	85	79	80	75	79	74	70	77	65	71	50	36	50	56	45	46	61	59	62	65	45	47
CV1	86	81	82	85	90	89	76	68	73	72	76	78	85	91	88	100	90	83	83	80	74	75	70	82	81	72	72	64	65	50	35	48	61	45	49	65	64	66	71	46	46
CV2	88	84	84	87	90	89	77	71	72	73	74	80	84	87	89	90	100	85	82	84	79	79	72	80	78	72	76	68	70	49	39	47	55	43	50	62	59	64	65	49	45
CV3	80	80	80	84	83	81	85	79	80	78	83	84	86	83	83	83	85	100	81	83	84	74	80	76	75	70	71	64	68	53	32	47	55	47	48	62	62	63	71	47	42

Figure 4. Annotator agreement (all pairs). Left to right: annotator agreement with each other, the learned metric, handcrafted metrics, and VLMs. Annotators background: A-K – 3D modellers, Des[0-2] - designers, CV[0-3] - computer vision engineers. Best zoom-in.

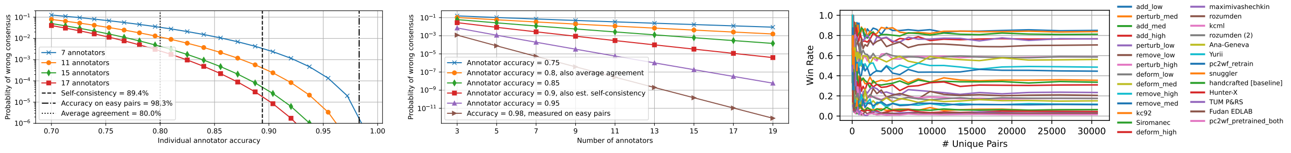


Figure 5. Probability of selecting wrong ”winner” depending on number of raters (left), individual accuracy (center), win rate (right).

because of wrong clicks. Average rater accuracy on these ”easy” pairs is 98.3%.

Self-consistency. We assessed intra-rater reliability by measuring how consistently annotators rated repeated pairs of wireframes, occasionally reversing pair order to mitigate order biases. These repeated evaluations constituted only a small subset of the total ratings (self-consistency checks are performed with 5% probability). The average self-consistency score is 89.4%, which could be partially attributed to the labeling mistakes and partially to the changing preferences during the annotation process, which we experienced ourselves.

Do we label enough? Under the assumption that there is some latent “correct” winner in any given pair, and N raters each independently select this “correct” winner with probability p, we can compute the probability that the majority is “wrong” (see Fig. 5). Therefore, the estimated panel error rate per pair is $\approx 1\%$ for 11 (expert) raters, and 0.25% for 17 (all) raters (assuming a significant 20% individual error rate). We also analyze the stability of our results and present adequacy analysis for the number of raters, comparisons, and houses. For each, we sweep a range of subsample sizes and resample 500 times at each size. For each subsample of a given size, we compute ranking implied by the win rates for that subset and rank correlation (Kendall τ) between the subset ranks and the rankings using the full dataset. We then construct a 95% Bootstrap CI across the 500 iterates for τ. The minimum number of raters/comparisons/houses needed for $\tau \geq 0.95$: comparisons ≥ 3350, houses ≥ 4, raters ≥ 8 (in all cases we have more than that).

Finding agreement. We compute an agreement score for each pair for annotators using the following simple rule: the same ranking gets 1 point, decisive ranking vs ”equal” gets 0.5 points, and the opposite ranking gets zero.

The agreement table between human annotators, metrics, and VLMs is shown in Fig. 4. The agreement table with all ”equal” rankings excluded, is shown in the supplementary.

Observation 4.1

When accounting for ties, there is a moderate global consensus among the human annotators; however, they form two distinct clusters that do not seem to depend on the annotator’s background. One group assigns more weights to the edge accuracy – correlated with edge F1 and Jaccard distance, and the second – vertex accuracy, correlated with corner F1 score.

The average agreement score is around 80% across all annotators, but within clusters, it increases to 85%. When annotators are decisive (select one of the reconstructions as clearly better rather than selecting ”equal”), then the average agreement increases to 91%, and no clusters are ob-

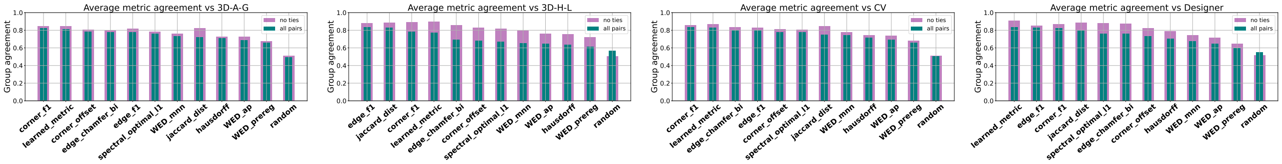


Figure 6. Metric ranking by agreement with group in average. Left to right: the first group of raters with more attention to vertices, the second group of raters with more attention to edges, computer vision engineers, and designers.

served. Agreement with the metrics (middle part of Fig. 4) suggests an explanation of preferences.

Observation 4.2

Human annotators pay more attention to correct parts of the reconstruction than the incorrect parts. Regardless of whether edges or vertices are considered, recall metrics agree more with human preferences than precision ones.

Cluster 1 (raters A-G, CV, Des1-2) mostly correlates with corner-based metrics, such as corner recall and corner F1-score, whereas Cluster 2 (raters H-K, Des0) correlates more with edge-based metrics, such as edge recall, edge F1-score and Jaccard distance. The second difference between clusters is that Cluster 2 is more likely to give an ”equal” score for the low-quality reconstruction, whereas Cluster 1 tried to rank reconstruction more decisively.

The metrics rankings w.r.t. different groups of raters is shown in Fig. 8.

Observation 4.3

The average agreement with human preferences of the top handcrafted metrics does not vary significantly. WED-based scores correlate with annotators the least.

The supplementary material shows the full agreement table. Furthermore, the ranking setup changes the human preferences for different metrics; for example, Jaccard distance performs much better for the decisive pairs compared to all other metrics. WED with pre-registration, which was used in the $S^2 3DR$ challenge, shows the worst correlation and is just slightly better than random chance. Other flavors of the WED metric, namely WED_mnn and WED_AP (used in Building3D Challenge), perform better but still worse than the rest. Graph structure metrics are in the middle. We also consider the agreement with VLMs. Despite initial success with individual examples, in our tests they did not perform meaningfully better than chance.

Observation 4.4

VLMs do not show significant agreement with human preferences in the wireframe ranking. The only exceptions are OpenAI models, as well as Grok2; yet those are only slightly better than chance. VLMs agree the most with the WED metrics family.

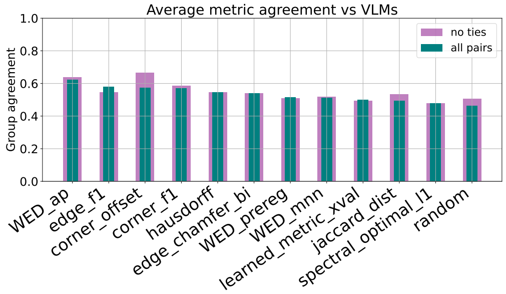


Figure 7. Metric ranking by agreement with VLMs

Finding the best reconstruction. To determine if there is a single ”quality” factor which explains the human expert judgments, we employ three distinct approaches to map these pairwise comparisons to a single ranking of the methods: simple win rates, a Bradley-Terry probability model, and a factor analysis based approach. While the methods differ, they point to highly concordant conclusions.

Observation 4.5

Human raters tend to rank equally all solutions that are below some quality threshold. Having no solution often ranks better, compared to a totally wrong reconstruction.

Simple Win Rate The first approach is similar to chess scoring – first a win-count table is computed. For each rated pair, each method receives 1 point for a win, 0.5 for a tie, and 0 for a loss. This scoring does not break ties but distributes the points evenly. Results with selected reconstruction methods are shown in Fig. 8. Consistent with the metric-human agreements, recall plays a more important role, and the wireframes with perfect recall – ”add_low” and ”add_med” are among the leaders in all groups. In other words, extra (erroneous) edges are considered less of an issue when compared to missing an edge or a vertex – the ”remove_*” family. One possible explanation resides in the information contained in the reconstruction: if all correct edges are present, we may identify and remove any erroneous ones. However, in many cases, inferring the exact position of an edge or vertex is not possible given the reconstruction alone if key information is missing.

The next set of solutions contains roughly correct but slightly noisy reconstructions - ”perturb_*” and ”deform*” family. One of the best solutions from the S23DR Challenge ranked better than highly deformed ground truth but worse than the less invasive corruptions.

The rest of the reconstructions get almost equal scores due to the high proportion of draws and lack of wins among themselves.

Bradley-Terry Model We have also modeled the quality of each solution using a Bradley-Terry (BT) [7, 11, 14, 35] preference model on the expressed preferences of the annotators (using the BT-Abilities as scores). The Bradley-Terry model defines the probability that item i is preferred over item j as:

$$P(i > j) = \frac{a_i}{a_i + a_j} \,, \tag{4}$$

where a_i, a_j are positive real numbers representing the latent strength of each item.

Following standard practice, we reparameterize these latent strengths as exponentials of real-valued parameters θ_i, giving $a_i = e^{\theta_i}$. This yields:

$$P(i > j) = \frac{e^{\theta_i}}{e^{\theta_i} + e^{\theta_j}} = \frac{1}{1 + e^{-(\theta_i - \theta_j)}} = \sigma(\theta_i - \theta_j) \,, \tag{5}$$

where $\sigma(x)$ is the sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$.

This formulation is further generalized by introducing a scale parameter s and offset o:

$$p_{ij} = \sigma \left(\frac{\theta_i - \theta_j}{s} + o \right) \,. \tag{6}$$

With $s = 1$ and $o = 0$, this reduces to the standard Bradley-Terry formulation. Alternatively, setting $s = 400$ and $o = 800$ yields the Elo scoring system familiar to chess players.

To estimate the latent abilities θ, we initialize each θ_i by sampling from an independent Gaussian distribution and then iteratively minimize the expectation of the following binary cross-entropy loss using stochastic gradient descent (SGD) with the Adam optimizer:

$$\mathcal{L} = \mathbb{E}_{(i,j)} \left[-y \log(p_{ij}) - (1 - y) \log(1 - p_{ij}) \right] \,, \tag{7}$$

where $y = 1$ if item i was chosen over item j, and $y = 0$ otherwise.

Factor Analysis To investigate whether the data reflect a single underlying dimension of quality, we additionally peruse a factor analysis based approach. We form the methods-by-raters table M such that M_{kl} is the rate at which rater k chose method l when they saw it. We hypothesize the empirical log-odds of these win-rates (rate l wins according to k), $\eta = \log \frac{M}{1-M}$ possess a low-rank structure (rank one in the ideal case); this would indicate a single dominant factor ("quality") governing outcomes. We apply singular value decomposition (SVD) to factorize $\eta = U \Sigma V^T$ and extract the first left singular vector of η containing the estimated quality scores.

Method	Empirical Win Rate	Implied Win Rate (BT)	Implied Win Rate (Elo)	BT Ability	Elo Score	Quality Factor
add_low	0.89	0.89	0.89	2.79	1937	0.03
add_med	0.86	0.86	0.86	2.47	1769	0.02
perturb_med	0.85	0.85	0.85	2.33	1739	0.02
add_high	0.82	0.82	0.82	2.04	1604	-0.02
perturb_low	0.79	0.79	0.79	1.79	1510	-0.01
remove_low	0.79	0.78	0.79	1.71	1498	-0.02
perturb_high	0.67	0.67	0.68	0.83	1144	-0.09
deform_med	0.67	0.67	0.66	0.83	1107	-0.09
deform_low	0.66	0.66	0.66	0.78	1094	-0.10
remove_high	0.65	0.65	0.65	0.67	1077	-0.10
remove_med	0.63	0.64	0.64	0.60	1027	-0.11
kc92	0.51	0.51	0.51	-0.25	698	-0.18
Siromanec	0.50	0.50	0.49	-0.34	645	-0.19
deform_high	0.50	0.50	0.50	-0.34	669	-0.19
maximivashechkin	0.39	0.39	0.39	-1.01	386	-0.27
rozumden	0.38	0.38	0.38	-1.10	373	-0.28
kcml	0.35	0.35	0.35	-1.28	291	-0.26
rozumden	0.34	0.34	0.33	-1.35	236	-0.26
Ana-Geneva	0.32	0.32	0.32	-1.47	221	-0.25
pc2wf_retrain	0.29	0.29	0.30	-1.65	157	-0.23
Yurii	0.29	0.29	0.29	-1.63	146	-0.25
snuggler	0.25	0.25	0.24	-1.92	15	-0.25
baseline	0.25	0.25	0.25	-1.91	19	-0.25
Hunter-X	0.22	0.22	0.22	-2.10	-43	-0.25
TUM	0.22	0.22	0.22	-2.13	-48	-0.26
Fudan EDLAB	0.21	0.21	0.21	-2.18	-76	-0.26
pc2wf_pretrained	0.20	0.20	0.20	-2.28	-117	-0.25

Table 1. Win rates for selected wireframe, according to the pairwise human annotations and estimated Elo.

Observation 4.6

We find a Kendall correlation coefficient >0.7 between the rankings implied by SVD and those implied by BT. This lends additional evidence to the hypothesis that there is a true "quality" factor driving the raters' views. The result is shown in Table 1.

Metric properties and "unit tests". We define a range of properties a good metric should have, implement tests for these properties, and report the results for all metrics. We use a dataset of 128 ground truth wireframes, which we disturb or alter and check the behavior of each metric. We report the percentage of wireframes where the property was valid for a given metric. Results are shown in Table 2. Most of the metrics pass the triangle inequality test, the identity of indiscernibles, and symmetry. None of the metrics is perfectly monotonic w.r.t. wireframe changes, which makes sense; for example, the precision metrics are insensitive to the number of predicted vertices/edges as long as each predicted element is aligned with an element of the ground truth. Hausdorff and WED with pre-registration pass the fewest tests, which is in line with their poor performance w.r.t. human preferences. Conversely – corner F1, edge F1, and Jaccard perform well and align well with the preferences of annotators. The spectral L1 distance and Chamfer edge distance are exceptions – the spectral distance scores well on properties, but not human alignment, and Chamfer one – vice versa.

Additional considerations. After speaking to the human

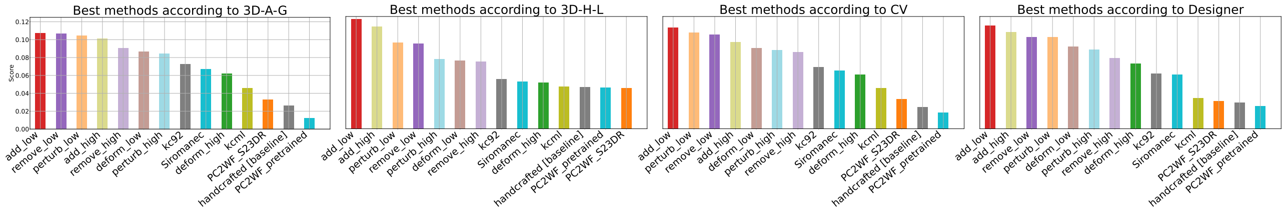


Figure 8. Methods scores according to different groups.

	Corner				Edge			WED				Spectral		IoU	Hausdorff	Chamfer
Test	Prec	Rec	F1	offset	Prec	Rec	F1	prereg	MNN	nearest	AP	L1	L2	Jaccard	dist.	edge
Monotonic																
Monotonic (wrong edges)	0.02	0.01	0.02	1.00	1.00	0.00	1.00	0.00	0.00	0.00	0.98	0.63	0.14	0.01	0.09	0.00
Monotonic (deform/split)	1.00	0.86	1.00	0.65	1.00	0.98	1.00	0.74	0.79	0.67	0.60	0.00	0.06	0.16	0.05	0.40
Monotonic (moving vertex)	1.00	1.00	1.00	0.98	1.00	1.00	1.00	0.99	1.00	0.99	1.00	0.99	0.99	0.45	0.99	1.00
Monotonic (disconnect edges)	0.52	0.00	0.50	0.31	0.00	0.00	0.00	0.26	1.00	1.00	1.00	0.00	0.01	1.00	0.00	0.02
Monotonic (delete vertices)	0.00	1.00	1.00	1.00	0.00	1.00	1.00	0.75	1.00	1.00	1.00	0.90	0.33	0.67	0.14	0.99
Monotonic (delete edges)	0.00	0.00	0.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.31	1.00	0.07	0.83
Identity																
Identity of indiscernibles	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.02	1.00	1.00	1.0	1.00	1.00	1.00	0.00	0.00
Near identity of indiscernibles	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.0	1.00	1.00	1.00	0.00	0.00
Symmetry																
Symmetry (0 mean, weighted)	1.00	1.00	1.00	0.88	1.00	1.00	1.00	0.44	0.44	0.44	0.44	0.44	0.90	1.00	0.84	1.00
Near symmetry (0 mean, weighted)	1.00	1.00	1.00	0.87	1.00	1.00	1.00	0.58	0.54	0.54	0.45	0.90	0.92	1.00	1.00	1.00
Symmetry (shift, weighted)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.88	0.99	0.92	0.98	0.99	1.00	1.00	0.82	1.00
Near symmetry (shift, weighted)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.92	0.99	1.00	1.00	1.00	1.00	1.00
Quasi-proportionality																
Quasi-proportionality (shift, far)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.72	0.05	0.00	0.00
Quasi-proportionality (shift, close)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.16	0.16	0.91	1.00	0.72	0.06	0.00	0.00
Triangle ineq																
Triangle ineq. (rand other)	0.99	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00	0.99	0.95	0.92	0.97	1.00	1.00	0.96
Triangle ineq. (add noise)	1.00	1.00	1.00	0.91	1.00	1.00	1.00	0.84	0.81	0.83	1.00	0.90	0.69	1.00	1.00	1.00
Triangle ineq. (del1/del2)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.90	0.98	0.98	1.00	1.00	0.66	1.00	1.00	0.43
Pass count	11/17	11/17	12/17	11/17	12/17	13/17	14/17	8/17	10/17	10/17	13/17	13/17	8/17	11/17	6/17	8/17

Table 2. Properties ”unit-test” results for metrics, percentage of tests passed. The test is passed if the result is $\geq 90\%$ (black).

annotators, we note the following. First, the 3D reconstruction experts are considering how the estimated reconstruction could help them in 3D modeling. There are two main approaches to using such help. The first (in case of noisy reconstructions) is to use the wireframe as a guide to create their own clean reconstruction. The second, if the reconstruction is already close enough, is to fix the wireframe to produce the final result. The Wireframe Edit Distance is designed to estimate how costly it would be to modify the predicted wireframe to match the ground truth. The issue with it is the set of operations – WED only considers vertex/edge deletion/insertion and vertex movement. In practice, one could fit a single edge to multiple noisy ones, bulk-delete a lot of wrong edges or vertices, and apply a rigid transform on the whole model. All of these operations are commonly used in 3D editing software, and WED could benefit from them.

Finally, if the wireframe is totally wrong, then it is better to have no reconstruction at all and start from scratch. In this ”low quality regime”, most human annotators see no difference between reconstructions if both are wrong.

Considering usage in competition and benchmarks, we would recommend the use of F1-score, despite the fact that recall-based metrics are more in line with human preferences, as it is less easy to game. For example, a dense grid of vertices and edges could score perfectly on recall but be useless in practice and score poorly on precision-based metrics.

5. Conclusion

We have studied how human preferences in structured reconstruction evaluation are explained via a wide range of metrics. We show that human preferences can be learned from a small number of examples by transferring from pretrained models which can subsequently be used to score unseen reconstructions. However, we conclude that additional study is warranted prior to relying on such learned metrics as the sole adjudication mechanism for competitions (especially those with strong incentives), because of the potential for reward hacking, gradient-based adversarial attacks, and the like. Based on our study, we recommend using a combination of edge-based (edge F1 or Jaccard score) and corner-based metrics (F1) for benchmarks and competitions. They better explain human preferences in ranking structured reconstructions than more complex and fragile graph-based metrics such as WED or spectral distances.

Acknowledgements We thank Sherwin Dale Sotto, Rosalie Cellacay, John Carlo Castro, Johsua Vernon Awog, Romeo Sapitanan, Francis Angel Viloria, Joenard Sarmiento, Danica B. Mercado, Zander Hechanova, Michael James Dulay, Joshua Muñoz, Jonah Franz R. De Leon, Viktor Rusov, Olha Mishkina, Yihe Wang and Tyler Nguyen for their help with wireframe annotations. We also thank Tolga Birdal, Caner Korkmaz, Hanzhi Chen, Daoyi Gao, and Ilke Demir for helpful discussions and dataset work at the early stage of the paper. Dmytro Mishkin is supported by MPO 60273/24/21300/21000 CEDMO 2.0 NPO project.

References

[1] 1st workshop on urban scene modeling: Where vision meets photogrammetry and graphics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024. 1

[2] Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, Teven Le Scao, Andy Lo, William Marshall, Louis Martin, Arthur Mensch, Pavankumar Muddireddy, Valera Nemychnikova, Marie Pellat, Patrick Von Platen, Nikhil Raghuraman, Baptiste Rozière, Alexandre Sablayrolles, Lucile Saulnier, Romain Sauvestre, Wendy Shang, Roman Soletskyi, Lawrence Stewart, Pierre Stock, Joachim Studnia, Sandeep Subramanian, Sagar Vaze, Thomas Wang, and Sophia Yang. Pixtral 12b, 2024. 4

[3] AI Anthropic. Claude 3.5 sonnet model card addendum. *Claude-3.5 Model Card*, 2024. 4

[4] Andre Araujo, Bingyi Cao, Cam Askew, Jack Sim, Maggie, Tobias Weyand, and Will Cukierski. Google landmark retrieval 2021. https://kaggle.com/competitions/landmark-retrieval-2021, 2021. Kaggle. 1

[5] Armen Avetisyan, Christopher Xie, Henry Howard-Jenkins, Tsun-Yi Yang, Samir Aroudj, Suvam Patra, Fuyang Zhang, Duncan Frost, Luke Holland, Campbell Orme, Jakob Engel, Edward Miller, Richard Newcombe, and Vasileios Balntas. Scenescript: Reconstructing scenes with an autoregressive structured language model. In *European Conference on Computer Vision (ECCV)*, 2024. 1

[6] Eric Brachmann, Jamie Wynn, Shuai Chen, Tommaso Cavallari, Áron Monszpart, Daniyar Turmukhambetov, and Victor Adrian Prisacariu. Scene coordinate reconstruction: Posing of image collections via incremental learning of a relocalizer. In *ECCV*, 2024. 2

[7] Ralph A. Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. 4, 7

[8] Li Cao, Yike Xu, Jianwei Guo, and Xiaoping Liu. Wireframenet: A novel method for wireframe generation from point cloud. *Computers and Graphics*, 115:226–235, 2023. 1

[9] Jiacheng Chen, Yiming Qian, and Yasutaka Furukawa. Heat: Holistic edge attention transformer for structured reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[10] Kseniya Cherenkova, Elona Dupont, Anis Kacem, Ilya Arzhannikov, Gleb Gusev, and Djamila Aouada. Sepicnet: Sharp edges recovery by parametric inference of curves in 3d shapes. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2727–2735, 2023. 2, 3

[11] Arpad E. Elo. *The Rating of Chessplayers, Past and Present*. Arco Publishing, 1978. Accessed: 2025-03-06. 7

[12] Gemini Team et al. Gemini: A family of highly capable multimodal models, 2024. 4

[13] OpenAI et al. Openai o1 system card, 2024. 4

[14] Mark E Glickman. Introductory note to 1928, 2013. 7

[15] Tomáš Hodaň, Martin Sundermeyer, Yann Labbé, Van Nguyen Nguyen, Gu Wang, Eric Brachmann, Bertram Drost, Vincent Lepetit, Carsten Rother, and Jiří Matas. BOP challenge 2023 on detection, segmentation and pose estimation of seen and unseen rigid objects. *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024. 1

[16] Perpetual Hope Akwensi, Akshay Bharadwaj, and Ruisheng Wang. Apc2mesh: Bridging the gap from occluded building façades to full 3d models. *ISPRS Journal of Photogrammetry and Remote Sensing*, 211:438–451, 2024. 2, 3

[17] Shangfeng Huang, Ruisheng Wang, Bo Guo, and Hongxin Yang. Pbwr: Parametric-building-wireframe reconstruction from aerial lidar point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27778–27787, 2024. 2, 3

[18] Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Kai Dang, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024. 4

[19] Apple Inc. Roomplan: Create 3d floor plans with iphone and ipad. https://developer.apple.com/augmented-reality/roomplan/, 2022. 1

[20] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image Matching across Wide Baselines: From Paper to Practice. *International Journal of Computer Vision*, 2020. 1, 2

[21] Matej Kristan, Jiri Matas, Aleš Leonardis, Tomas Vojir, Roman Pflugfelder, Gustavo Fernandez, Georg Nebehay, Fatih Porikli, and Luka Čehovin. A novel performance evaluation methodology for single-target trackers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(11):2137–2155, 2016. 1

[22] Jack Langerman, Caner Korkmaz, Hanzhi Chen, Daoyi Gao, Ilke Demir, Dmytro Mishkin, and Tolga Birdal. S23dr competition at 1st workshop on urban scene modeling @ cvpr 2024. https://huggingface.co/usm3d, 2024. 1, 2, 3

[23] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler. MOTChallenge 2015: Towards a benchmark for multi-target tracking. *arXiv:1504.01942 [cs]*, 2015. arXiv: 1504.01942. 1

[24] Yujia Liu, Stefano D'Aronco, Konrad Schindler, and Jan Dirk Wegner. Pc2wf: 3d wireframe reconstruction from raw point clouds. In *International Conference on Learning Representations*, 2021. 1, 2, 3

[25] Yicheng Luo, Jing Ren, Xuefei Zhe, Di Kang, Yajing Xu, Peter Wonka, and Linchao Bao. Lc2wf:learning to construct 3d building wireframes from 3d line clouds. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2022. 1

[26] Wenchao Ma, Bin Tan, Nan Xue, Tianfu Wu, Xianwei Zheng, and Gui-Song Xia. How-3d: Holistic 3d wireframe perception from a single image. In *International Conference on 3D Vision*, 2022. 1

[27] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[28] OpenRouter. Openrouter: Unified api for ai models, 2024. Accessed: 2025-03-07. 4

[29] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 4

[30] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *CVPR*, 2018. 1

[31] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. 1

[32] Alberto Sanfeliu and King-Sun Fu. A distance measure between attributed relational graphs for pattern recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(3):353–362, 1983. 3

[33] Ruisheng Wang, Shangfeng Huang, and Hongxin Yang. Building3D: An Urban-Scale Dataset and Benchmarks for Learning Roof Structures from Point Clouds . In *ICCV*, pages 20019–20029, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 1, 3

[34] xAI. Grok-2: Advancements in ai language modeling. 2024. Accessed: 2025-03-07. 4

[35] Ernst Zermelo. Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 29:436–460, 1929. 7

[36] Yichao Zhou, Haozhi Qi, and Yi Ma. End-to-end wireframe parsing. In *ICCV 2019*, 2019. 1

Explaining Human Preferences via Metrics for Structured 3D Reconstruction

Supplementary Material

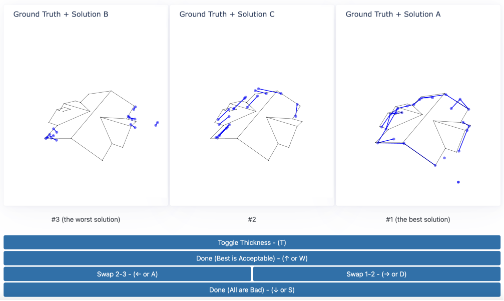

Ground Truth + Solution B
Ground Truth + Solution C
Ground Truth + Solution A
#3 (the worst solution)
#2
#1 (the best solution)
Toggle Thickness - (T)
Done (Best is Acceptable) - (↑ or W)
Swap 2-3 - (← or A)
Swap 1-2 - (→ or D)
Done (All are Bad) - (↓ or S)

Figure 9. First iteration wireframe ranking interface for human annotators

6. Ranking and Annotator workflow

Annotators were selected from the same pool of 3D modelers who created the ground truth wireframes. The experiment is conducted through their employer, and compensated at their usual fair hourly rate.

Most raters rate a few more pairs than minimally required 3500 because of the self-consistency checks. In prior experiments, we observed a drop-off in self-consistency after around 400 pairs. Therefore, we inserted a following pop-up every 350 pairs: “Time for a Break! Taking regular breaks helps maintain rating quality. We recommend: Stand up and stretch, Look away from the screen, Take a short walk if possible.” We found this to be effective in maintaining rating quality. One rater completed all ratings (over 3500) in under 7 hours. All other raters spread the task over two or more (up to 27) days. Mean and median ratings per day are 1327 and 207, respectively.

7. Ranking triplets

We have experimented with different forms of rankind the reconstructions, one of which is shown in Figure 9. The annotators were sorting triplets, and also marking if the best solution is acceptable. However, it took people much longer time to process one triplet, and in addition to that, the ranks 2-3 was much less reliable, than 1-2. In the end, we opted for simplicity and also added a message asking annotators to take a break after each 350 pairs.

8. Length Weighted Spectral Graph Distances

incorporate both topological and geometric information by framing graph (wireframe) distance in terms of distances between the spectra of weighted graph Laplacians. We measure the spectral distance using the 2-Wasserstein metric between the eigenvalue distributions:

$$SD(G_1, G_2) := W_2(\lambda(L_1), \lambda(L_2)) \tag{8}$$

where $\lambda(L)$ denotes the spectrum of the Laplacian L.

For a graph $G = (V, E)$, the weighted graph Laplacian is defined:

$$L := D - A \tag{9}$$

where D is the weighted degree matrix ($|V| \times |V|$ diagonal matrix with each diagonal entry containing the sum of the lengths of edges incident to that vertex), and A is the weighted adjacency matrix ($|V| \times |V|$ with $A_{ij} = \|V_i - V_j\|_2$ iff $(i, j) \in E$ and 0 otherwise).

9. Full agreement tables

The full annotator-metric-VLM agreement table is shown in Figures 11, 12, and the metric rankings are shown in Figure 10.

10. VLM Prompts

All the VLMs are prompted with the following text.

```
Here we see two possible wireframe
reconstructions of houses (shown in blue)
superimposed on top of the ground truth
wireframe (shown in black).
Please describe the quality of each of
the two reconstructions (Left and Right).
If you don’t see any blue lines it is
because the reconstruction is incomplete.
Which reconstruction most closely matches
the ground truth
(end by printing the final answer in all
caps:  "LEFT" or "RIGHT")?
```

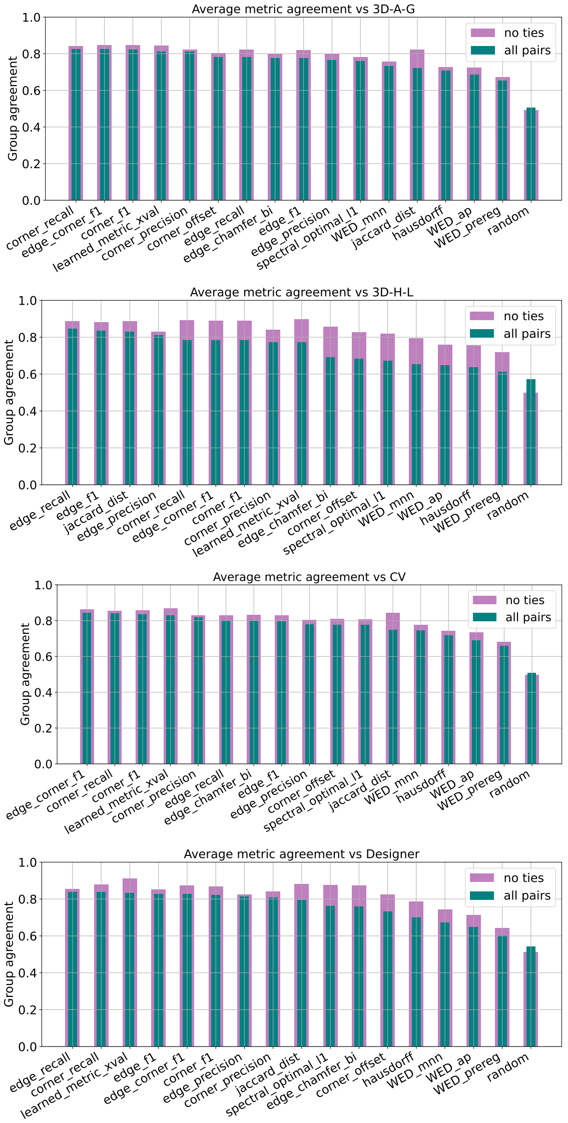
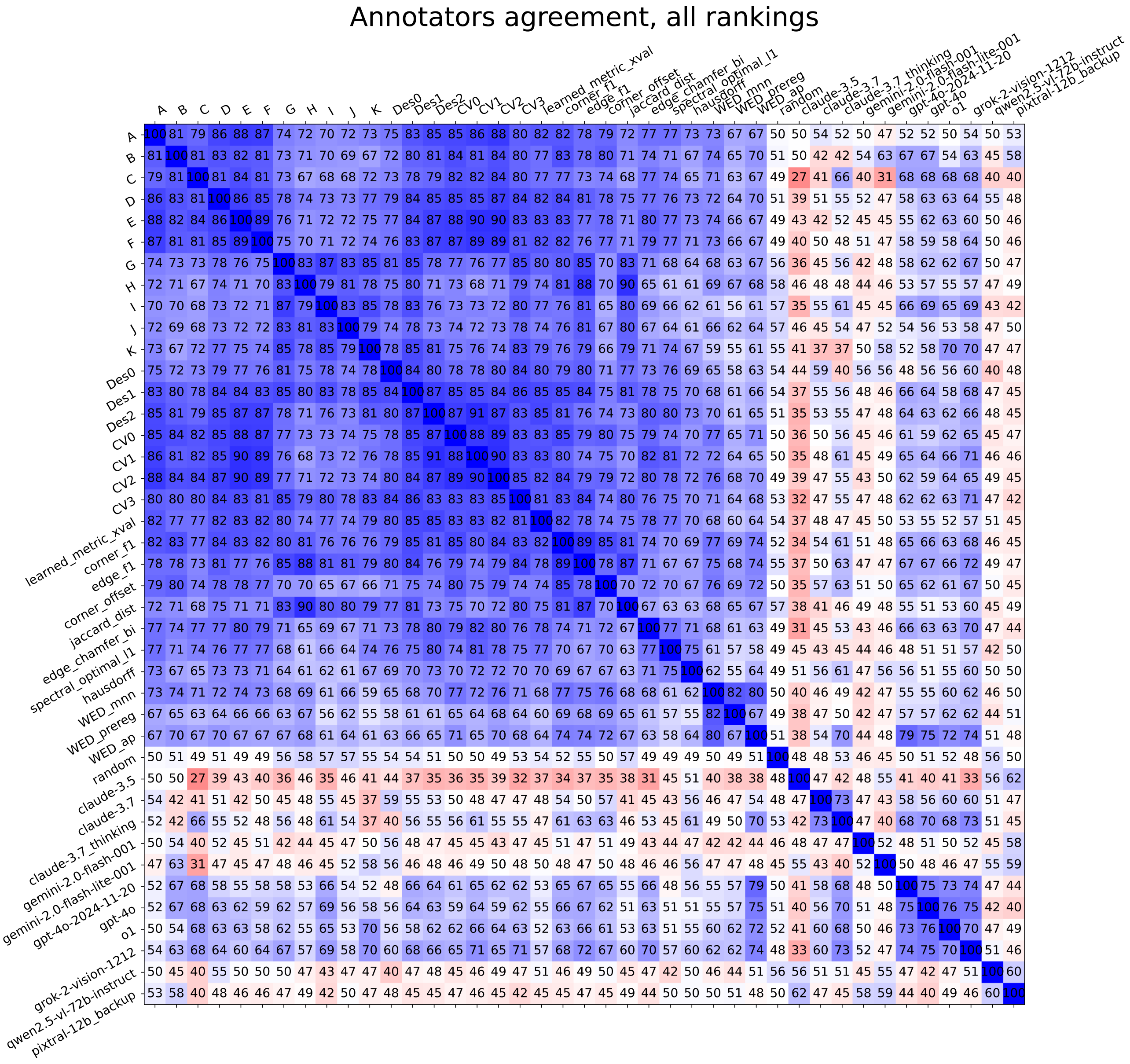


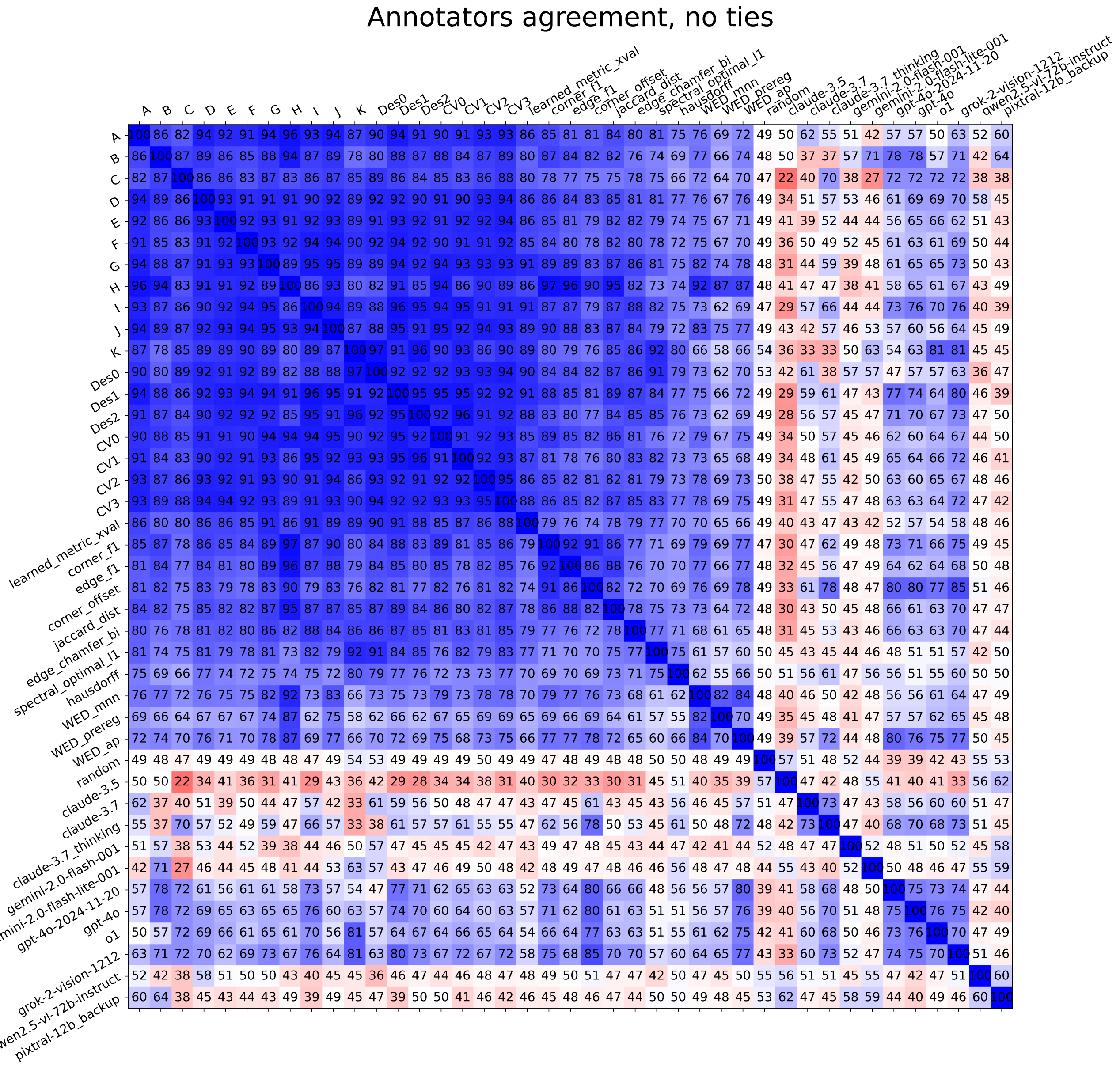
Figure 10. Metric ranking by agreement with group in average. From top to bottom: group of raters 1 with more attention to vertices, group of raters 2 with more attention to edges, computer vision engineers

Annotators agreement, all rankings



	A	B	C	D	E	F	G	H	I	J	K	Des0	Des1	Des2	CV0	CV1	CV2	CV3	learned_metric_xval	corner_f1	edge_f1	corner_offset	jaccard_dist	edge_chamfer_bi	spectral_optimal_l1	hausdorff	WED_mnn	WED_prereg	WED_ap	random	claude-3.5	claude-3.7	claude-3.7_thinking	gemini-2.0-flash-001	gemini-2.0-flash-lite-001	gpt-4o-2024-11-20	gpt-4o	o1	grok-2-vision-1212	qwen2.5-vl-72b-instruct	pixtral-12b_backup
A	100	81	79	86	88	87	74	72	70	72	73	75	83	85	85	86	88	80	82	82	78	79	72	77	77	73	73	67	67	50	50	54	52	50	47	52	52	50	54	50	53
B	81	100	81	83	82	81	73	71	70	69	67	72	80	81	84	81	84	80	77	83	78	80	71	74	71	67	74	65	70	51	50	42	42	54	63	67	67	54	63	45	58
C	79	81	100	81	84	81	73	67	68	68	72	73	78	79	82	82	84	80	77	77	73	74	68	77	74	65	71	63	67	49	27	41	66	40	31	68	68	68	68	40	40
D	86	83	81	100	86	85	78	74	73	73	77	79	84	85	85	85	87	84	82	84	81	78	75	77	76	73	72	64	70	51	39	51	55	52	47	58	63	63	64	55	48
E	88	82	84	86	100	89	76	71	72	72	75	77	84	87	88	90	90	83	83	83	77	78	71	80	77	73	74	66	67	49	43	42	52	45	45	55	62	63	60	50	46
F	87	81	81	85	89	100	75	70	71	72	74	76	83	87	87	89	89	81	82	82	76	77	71	79	77	71	73	66	67	49	40	50	48	51	47	58	59	58	64	50	46
G	74	73	73	78	76	75	100	83	87	83	85	81	85	78	77	76	77	85	80	80	85	70	83	71	68	64	68	63	67	56	36	45	56	42	48	58	62	62	67	50	47
H	72	71	67	74	71	70	83	100	79	81	78	75	80	71	73	68	71	79	74	81	88	70	90	65	61	61	69	67	68	58	46	48	48	44	46	53	57	55	57	47	49
I	70	70	68	73	72	71	87	79	100	83	85	78	83	76	73	73	72	80	77	76	81	65	80	69	66	62	61	56	61	57	35	55	61	45	45	66	69	65	69	43	42
J	72	69	68	73	72	72	83	81	83	100	79	74	78	73	74	72	73	78	74	76	81	67	80	67	64	61	66	62	64	57	46	45	54	47	52	54	56	53	58	47	50
K	73	67	72	77	75	74	85	78	85	79	100	78	85	81	75	76	74	83	79	76	79	66	79	71	74	67	59	55	61	55	41	37	37	50	58	52	58	70	70	47	47
Des0	75	72	73	79	77	76	81	75	78	74	78	100	84	80	78	78	80	84	80	79	80	71	77	73	76	69	65	58	63	54	44	59	40	56	56	48	56	56	60	40	48
Des1	83	80	78	84	84	83	85	80	83	78	85	84	100	87	85	85	84	86	85	85	84	75	81	78	75	70	68	61	66	54	37	55	56	48	46	66	64	58	68	47	45
Des2	85	81	79	85	87	87	78	71	76	73	81	80	87	100	87	91	87	83	85	81	76	74	73	80	80	73	70	61	65	51	35	53	55	47	48	64	63	62	66	48	45
CV0	85	84	82	85	88	87	77	73	73	74	75	78	85	87	100	88	89	83	83	85	79	80	75	79	74	70	77	65	71	50	36	50	56	45	46	61	59	62	65	45	47
CV1	86	81	82	85	90	89	76	68	73	72	76	78	85	91	88	100	90	83	83	80	74	75	70	82	81	72	72	64	65	50	35	48	61	45	49	65	64	66	71	46	46
CV2	88	84	84	87	90	89	77	71	72	73	74	80	84	87	89	90	100	85	82	84	79	79	72	80	78	72	76	68	70	49	39	47	55	43	50	62	59	64	65	49	45
CV3	80	80	80	84	83	81	85	79	80	78	83	84	86	83	83	83	85	100	81	83	84	74	80	76	75	70	71	64	68	53	32	47	55	47	48	62	62	63	71	47	42
learned_metric_xval	82	77	77	82	83	82	80	74	77	74	79	80	85	85	83	83	82	81	100	82	78	74	75	78	77	70	68	60	64	54	37	48	47	45	50	53	55	52	57	51	45
corner_f1	82	83	77	84	83	82	80	81	76	76	76	79	85	81	85	80	84	83	82	100	89	85	81	74	70	69	77	69	74	52	34	54	61	51	48	65	66	63	68	46	45
edge_f1	78	78	73	81	77	76	85	88	81	81	79	80	84	76	79	74	79	84	78	89	100	78	87	71	67	67	75	68	74	55	37	50	63	47	47	67	67	66	72	49	47
corner_offset	79	80	74	78	78	77	70	70	65	67	66	71	75	74	80	75	79	74	74	85	78	100	70	72	70	67	76	69	72	50	35	57	63	51	50	65	62	61	67	50	45
jaccard_dist	72	71	68	75	71	71	83	90	80	80	79	77	81	73	75	70	72	80	75	81	87	70	100	67	63	63	68	65	67	57	38	41	46	49	48	55	51	53	60	45	49
edge_chamfer_bi	77	74	77	77	80	79	71	65	69	67	71	73	78	80	79	82	80	76	78	74	71	72	67	100	77	71	68	61	63	49	31	45	53	43	46	66	63	63	70	47	44
spectral_optimal_l1	77	71	74	76	77	77	68	61	66	64	74	76	75	80	74	81	78	75	77	70	67	70	63	77	100	75	61	57	58	49	45	43	45	44	46	48	51	51	57	42	50
hausdorff	73	67	65	73	73	71	64	61	62	61	67	69	70	73	70	72	72	70	70	69	67	67	63	71	75	100	62	55	64	49	51	56	61	47	56	56	51	55	60	50	50
WED_mnn	73	74	71	72	74	73	68	69	61	66	59	65	68	70	77	72	76	71	68	77	75	76	68	68	61	62	100	82	80	50	40	46	49	42	47	55	55	60	62	46	50
WED_prereg	67	65	63	64	66	66	63	67	56	62	55	58	61	61	65	64	68	64	60	69	68	69	65	61	57	55	82	100	67	49	38	47	50	42	47	57	57	62	62	44	51
WED_ap	67	70	67	70	67	67	67	68	61	64	61	63	66	65	71	65	70	68	64	74	74	72	67	63	58	64	80	67	100	51	38	54	70	44	48	79	75	72	74	51	48
random	50	51	49	51	49	49	56	58	57	57	55	54	54	51	50	50	49	53	54	52	55	50	57	49	49	49	50	49	51	100	48	48	53	46	45	50	51	52	48	56	50
claude-3.5	50	50	27	39	43	40	36	46	35	46	41	44	37	35	36	35	39	32	37	34	37	35	38	31	45	51	40	38	38	48	100	47	42	48	55	41	40	41	33	56	62
claude-3.7	54	42	41	51	42	50	45	48	55	45	37	59	55	53	50	48	47	47	48	54	50	57	41	45	43	56	46	47	54	48	47	100	73	47	43	58	56	60	60	51	47
claude-3.7_thinking	52	42	66	55	52	48	56	48	61	54	37	40	56	55	56	61	55	55	47	61	63	63	46	53	45	61	49	50	70	53	42	73	100	47	40	68	70	68	73	51	45
gemini-2.0-flash-001	50	54	40	52	45	51	42	44	45	47	50	56	48	47	45	45	43	47	45	51	47	51	49	43	44	47	42	42	44	46	48	47	47	100	52	48	51	50	52	45	58
gemini-2.0-flash-lite-001	47	63	31	47	45	47	48	46	45	52	58	56	46	48	46	49	50	48	50	48	47	50	48	46	46	56	47	47	48	45	55	43	40	52	100	50	48	46	47	55	59
gpt-4o-2024-11-20	52	67	68	58	55	58	58	53	66	54	52	48	66	64	61	65	62	62	53	65	67	65	55	66	48	56	55	57	79	50	41	58	68	48	50	100	75	73	74	47	44
gpt-4o	52	67	68	63	62	59	62	57	69	56	58	56	64	63	59	64	59	62	55	66	67	62	51	63	51	51	55	57	75	51	40	56	70	51	48	75	100	76	75	42	40
o1	50	54	68	63	63	58	62	55	65	53	70	56	58	62	62	66	64	63	52	63	66	61	53	63	51	55	60	62	72	52	41	60	68	50	46	73	76	100	70	47	49
grok-2-vision-1212	54	63	68	64	60	64	67	57	69	58	70	60	68	66	65	71	65	71	57	68	72	67	60	70	57	60	62	62	74	48	33	60	73	52	47	74	75	70	100	51	46
qwen2.5-vl-72b-instruct	50	45	40	55	50	50	50	47	43	47	47	40	47	48	45	46	49	47	51	46	49	50	45	47	42	50	46	44	51	56	56	51	51	45	55	47	42	47	51	100	60
pixtral-12b_backup	53	58	40	48	46	46	47	49	42	50	47	48	45	45	47	46	45	42	45	45	47	45	49	44	50	50	50	51	48	50	62	47	45	58	59	44	40	49	46	60	100

Figure 11. Annotator-metrics-LLM agreement (all pairs). Annotators background: A-K – 3D modellers, Des[0-2] - designers, CV[0-3] - computer vision engineers. Best zoom-in.



	A	B	C	D	E	F	G	H	I	J	K	Des0	Des1	Des2	CV0	CV1	CV2	CV3	learned_metric_xval	corner_f1	edge_f1	corner_offset	jaccard_dist	edge_chamfer_bi	spectral_optimal_l1	hausdorff	WED_mnn	WED_prereg	WED_ap	random	claude-3.5	claude-3.7	claude-3.7_thinking	gemini-2.0-flash-001	gemini-2.0-flash-lite-001	gpt-4o-2024-11-20	gpt-4o	o1	grok-2-vision-1212	qwen2.5-vl-72b-instruct	pixtral-12b_backup
A	100	86	82	94	92	91	94	96	93	94	87	90	94	91	90	91	93	93	86	85	81	81	84	80	81	75	76	69	72	49	50	62	55	51	42	57	57	50	63	52	60
B	86	100	87	89	86	85	88	94	87	89	78	80	88	87	88	84	87	89	80	87	84	82	82	76	74	69	77	66	74	48	50	37	37	57	71	78	78	57	71	42	64
C	82	87	100	86	86	83	87	83	86	87	85	89	86	84	85	83	86	88	80	78	77	75	75	78	75	66	72	64	70	47	22	40	70	38	27	72	72	72	72	38	38
D	94	89	86	100	93	91	91	91	90	92	89	92	92	90	91	90	93	94	86	86	84	83	85	81	81	77	76	67	76	49	34	51	57	53	46	61	69	69	70	58	45
E	92	86	86	93	100	92	93	91	92	93	89	91	93	92	91	92	92	94	86	85	81	79	82	82	79	74	75	67	71	49	41	39	52	44	44	56	65	66	62	51	43
F	91	85	83	91	92	100	93	92	94	94	90	92	94	92	90	91	91	92	85	84	80	78	82	80	78	72	75	67	70	49	36	50	49	52	45	61	63	61	69	50	44
G	94	88	87	91	93	93	100	89	95	95	89	89	94	92	94	93	93	93	91	89	89	83	87	86	81	75	82	74	78	48	31	44	59	39	48	61	65	65	73	50	43
H	96	94	83	91	91	92	89	100	86	93	80	82	91	85	94	86	90	89	86	97	96	90	95	82	73	74	92	87	87	48	41	47	47	38	41	58	65	61	67	43	49
I	93	87	86	90	92	94	95	86	100	94	89	88	96	95	94	95	91	91	91	87	87	79	87	88	82	75	73	62	69	47	29	57	66	44	44	73	76	70	76	40	39
J	94	89	87	92	93	94	95	93	94	100	87	88	95	91	95	92	94	93	89	90	88	83	87	84	79	72	83	75	77	49	43	42	57	46	53	57	60	56	64	45	49
K	87	78	85	89	89	90	89	80	89	87	100	97	91	96	90	93	86	90	89	80	79	76	85	86	92	80	66	58	66	54	36	33	33	50	63	54	63	81	81	45	45
Des0	90	80	89	92	91	92	89	82	88	88	97	100	92	92	92	93	93	94	90	84	84	82	87	86	91	79	73	62	70	53	42	61	38	57	57	47	57	57	63	36	47
Des1	94	88	86	92	93	94	94	91	96	95	91	92	100	95	95	95	92	92	91	88	85	81	89	87	84	77	75	66	72	49	29	59	61	47	43	77	74	64	80	46	39
Des2	91	87	84	90	92	92	92	85	95	91	96	92	95	100	92	96	91	92	88	83	80	77	84	85	85	76	73	62	69	49	28	56	57	45	47	71	70	67	73	47	50
CV0	90	88	85	91	91	90	94	94	94	95	90	92	95	92	100	91	92	93	85	89	85	82	86	81	76	72	79	67	75	49	34	50	57	45	46	62	60	64	67	44	50
CV1	91	84	83	90	92	91	93	86	95	92	93	93	95	96	91	100	92	93	87	81	78	76	80	83	82	73	73	65	68	49	34	48	61	45	49	65	64	66	72	46	41
CV2	93	87	86	93	92	91	93	90	91	94	86	93	92	91	92	92	100	95	86	85	82	81	82	81	79	73	78	69	73	50	38	47	55	42	50	63	60	65	67	48	46
CV3	93	89	88	94	94	92	93	89	91	93	90	94	92	92	93	93	95	100	88	86	85	82	87	85	83	77	78	69	75	49	31	47	55	47	48	63	63	64	72	47	42
learned_metric_xval	86	80	80	86	86	85	91	86	91	89	89	90	91	88	85	87	86	88	100	79	76	74	78	79	77	70	70	65	66	49	40	43	47	43	42	52	57	54	58	48	46
corner_f1	85	87	78	86	85	84	89	97	87	90	80	84	88	83	89	81	85	86	79	100	92	91	86	77	71	69	79	69	77	47	30	47	62	49	48	73	71	66	75	49	45
edge_f1	81	84	77	84	81	80	89	96	87	88	79	84	85	80	85	78	82	85	76	92	100	86	88	76	70	70	77	66	77	48	32	45	56	47	49	64	62	64	68	50	48
corner_offset	81	82	75	83	79	78	83	90	79	83	76	82	81	77	82	76	81	82	74	91	86	100	82	72	70	69	76	69	78	49	33	61	78	48	47	80	80	77	85	51	46
jaccard_dist	84	82	75	82	82	82	87	95	87	87	85	87	89	84	86	80	82	87	78	86	88	82	100	78	75	73	73	64	72	48	30	43	50	45	48	66	61	63	70	47	47
edge_chamfer_bi	80	76	78	81	82	80	86	82	88	84	86	86	87	85	81	83	81	85	79	77	76	72	78	100	77	71	68	61	65	48	31	45	53	43	46	66	63	63	70	47	44
spectral_optimal_l1	81	74	75	81	79	78	81	73	82	79	92	91	84	85	76	82	79	83	77	71	70	70	75	77	100	75	61	57	60	50	45	43	45	44	46	48	51	51	57	42	50
hausdorff	75	69	66	77	74	72	75	74	75	72	80	79	77	76	72	73	73	77	70	69	70	69	73	71	75	100	62	55	66	50	51	56	61	47	56	56	51	55	60	50	50
WED_mnn	76	77	72	76	75	75	82	92	73	83	66	73	75	73	79	73	78	78	70	79	77	76	73	68	61	62	100	82	84	48	40	46	50	42	48	56	56	61	64	47	49
WED_prereg	69	66	64	67	67	67	74	87	62	75	58	62	66	62	67	65	69	69	65	69	66	69	64	61	57	55	82	100	70	49	35	45	48	41	47	57	57	62	65	45	48
WED_ap	72	74	70	76	71	70	78	87	69	77	66	70	72	69	75	68	73	75	66	77	77	78	72	65	60	66	84	70	100	49	39	57	72	44	48	80	76	75	77	50	45
random	49	48	47	49	49	49	48	48	47	49	54	53	49	49	49	49	50	49	49	47	48	49	48	48	50	50	48	49	49	100	57	51	48	52	44	39	39	42	43	55	53
claude-3.5	50	50	22	34	41	36	31	41	29	43	36	42	29	28	34	34	38	31	40	30	32	33	30	31	45	51	40	35	39	57	100	47	42	48	55	41	40	41	33	56	62
claude-3.7	62	37	40	51	39	50	44	47	57	42	33	61	59	56	50	48	47	47	43	47	45	61	43	45	43	56	46	45	57	51	47	100	73	47	43	58	56	60	60	51	47
claude-3.7_thinking	55	37	70	57	52	49	59	47	66	57	33	38	61	57	57	61	55	55	47	62	56	78	50	53	45	61	50	48	72	48	42	73	100	47	40	68	70	68	73	51	45
gemini-2.0-flash-001	51	57	38	53	44	52	39	38	44	46	50	57	47	45	45	45	42	47	43	49	47	48	45	43	44	47	42	41	44	52	48	47	47	100	52	48	51	50	52	45	58
gemini-2.0-flash-lite-001	42	71	27	46	44	45	48	41	44	53	63	57	43	47	46	49	50	48	42	48	49	47	48	46	46	56	48	47	48	44	55	43	40	52	100	50	48	46	47	55	59
gpt-4o-2024-11-20	57	78	72	61	56	61	61	58	73	57	54	47	77	71	62	65	63	63	52	73	64	80	66	66	48	56	56	57	80	39	41	58	68	48	50	100	75	73	74	47	44
gpt-4o	57	78	72	69	65	63	65	65	76	60	63	57	74	70	60	64	60	63	57	71	62	80	61	63	51	51	56	57	76	39	40	56	70	51	48	75	100	76	75	42	40
o1	50	57	72	69	66	61	65	61	70	56	81	57	64	67	64	66	65	64	54	66	64	77	63	63	51	55	61	62	75	42	41	60	68	50	46	73	76	100	70	47	49
grok-2-vision-1212	63	71	72	70	62	69	73	67	76	64	81	63	80	73	67	72	67	72	58	75	68	85	70	70	57	60	64	65	77	43	33	60	73	52	47	74	75	70	100	51	46
qwen2.5-vl-72b-instruct	52	42	38	58	51	50	50	43	40	45	45	36	46	47	44	46	48	47	48	49	50	51	47	47	42	50	47	45	50	55	56	51	51	45	55	47	42	47	51	100	60
pixtral-12b_backup	60	64	38	45	43	44	43	49	39	49	45	47	39	50	50	41	46	42	46	45	48	46	47	44	50	50	49	48	45	53	62	47	45	58	59	44	40	49	46	60	100

Figure 12. Annotator-metrics-LLM agreement (excluding ties pairs). annotators agreement to each other, handcrafted metrics, and visual language models. Annotators background: A-K – 3D modellers, Des[0-2] - designers, CV[0-3] - computer vision engineers. Best zoom-in.