KαLOS finds Consensus: A Meta-Algorithm for Evaluating Inter-Annotator Agreement in Complex Vision Tasks

Supplementary Material

7. Foundations

7.1. Recap of Krippendorff's Alpha

Goal. To quantify inter-annotator agreement using a metric that is robust to missing data, variable numbers of raters, and small sample sizes. Unlike simple percentage agreement, K-α corrects for chance, ensuring that the reported score reflects genuine consensus rather than random coincidence.

Problem. Standard correlation metrics require fully paired data where every rater annotates every item. In complex vision tasks, this assumption rarely holds. We require a method that distinguishes between a rater who was *not assigned* a task (implicit absence) and a rater who *saw* the image but found nothing (explicit absence/active disagreement).

Solution. We utilize the nominal form of K-α. The calculation proceeds in two steps: transforming the reliability matrix into a coincidence matrix, and computing the ratio of observed to expected disagreement.

1. The Coincidence Matrix. The pipeline first flattens the reliability matrix (columns=units u, rows=raters r) into a coincidence matrix. For every unit u, we generate all possible unique pairs of labels provided by the assigned raters. The handling of empty cells in the reliability matrix is the critical differentiator:

- **Implicit Absence (Missing Data):** If a rater provides no label because they were never asked to annotate that specific image or class, the entry is marked `NAN`. These entries generate no pairs and do not influence the score. This natively handles:
  - *Sparse assignment:* Images annotated by different subsets of raters.
  - *Open World / Class Subsets:* Scenarios where specific raters are assigned only specific class subsets (e.g., rater A labels "animals", rater B labels "vehicles"). If Rater A does not annotate a "car", it is missing data, not disagreement.
- **Explicit Absence (Active Disagreement):** If a rater is assigned to the image/class but provides no annotation for a discovered unit u, this is treated as the class `NO_OBJECT`. A disagreement between `Class A` and `NO_OBJECT` is treated mathematically identical to a disagreement between `Class A` and `Class B`.

2. The Metric (K-α). We compute the agreement α by comparing the observed coincidence of labels against the distribution expected by chance.

$$\alpha = 1 - \frac{D_o}{D_e} = \frac{(n-1)\sum_c o_{cc} - \sum_c n_c(n_c-1)}{n(n-1) - \sum_c n_c(n_c-1)} \quad (10)$$

Where:

- n: The total number of values paired across all units in the coincidence matrix.
- n_c: The total count of class c (frequency) across the entire dataset.
- o_{cc}: The observed count of pairs where both raters agreed on class c (the diagonal of the coincidence matrix).
- D_o: Observed disagreement (frequency of mismatched pairs).
- D_e: Expected disagreement (frequency of mismatches if raters assigned labels randomly based on the population distribution n_c).

Intuition. The metric operates as a ratio of signal to noise. If raters agree perfectly, the observed matches (o_{cc}) maximize, resulting in $\alpha = 1$. If raters agree only as often as a randomized shuffle of their collective labels would predict, $D_o \approx D_e$, resulting in $\alpha = 0$. Systemic disagreement (worse than chance) yields $\alpha < 0$.

7.2. Recap: Distributional Validation

In complex annotation tasks, choosing a distance function d (e.g., IoU, L2, OKS) is often heuristic. Braylan *et al.* [5] propose eliminating this arbitrariness by validating d against the dataset's inherent statistical properties. They posit that a valid distance function must maximize the separation between two distributions:

1. Observed Disagreement (D_o): The pairwise distances between different raters annotating the same item (representing signal).
2. Expected Disagreement (D_e): The pairwise distances between raters annotating different items (representing chance/noise).

If d is valid, D_o should be stochastically smaller than D_e. This separation is quantified using the Kolmogorov-Smirnov (KS) statistic:

$$KS = \sup_x |F_{D_o}(x) - F_{D_e}(x)| \quad (11)$$

where F is the empirical cumulative distribution function. In KαLOS, we adopt this logic not to replace α, but as a calibration step (Sec. 3.3) to empirically select the task-optimal d_{loc} and the transition threshold τ^* where signal becomes distinguishable from noise.

Table 1. Overview of datasets with repeated annotations. Bold names are the ones used for our main analysis.

Dataset	Task(s)	Amount of Raters per Image/Scene	Annotator Type	Rater Identification + Intra-Rater Evaluation	Guideline and Annotation Process	Automated Proposals
TexBiG [41]	Instance Segmentation	3–4 (train), 5 (test)	Mixed (Experts + Non-Experts)	Yes + No	Same	no
VinDr-CXR [32]	Object Detection	3 (train), 5 (test)	Experts (Radiologists)	Yes + No	Same	no
LVIS Consistency Subset [16]	Instance Segmentation	2	Non-Experts	No + No	Same	no
COCO Reannotated [27]	Object Detection	2	Non-Experts	No + No	Different	yes
Open Images Reannotated [27]	Object Detection	2	Non-Experts	No + No	Different	yes
NuCLS [2]	Instance Segmentation	11–39	Mixed (Experts + Non-Experts)	Yes + Yes	Same	yes
LIDC-IDRI (TCIA) [3]	Voxel Grid / 3D Volume	4	Experts (Radiologists)	No + No	Different (see text)	no
MARS [38]	Video Pose Estimation and Action Detection (behavior)	5 (pose), 8 (behavior; no localization)	Non-Experts (pose) + Experts (behavior)	Yes + Yes	Same	no

7.3. Notations

Ground Truth

- $y_{ij} = (b_{ij}, c_{ij})$: The unknown true label for instance j in image i.
- $Y_i = y_{ij}{}_{j=1}^{N_i}$: The set of all true instances in image i.

Annotations (Observed Data)

- $\tilde{y}_{ik}^r = (\tilde{b}_{ik}^r, \tilde{c}_{ik}^r)$: An annotation from rater r.
- $\tilde{Y}_i^r$: The set of all annotations from rater r for image i.
- r: A specific rater.
- R: The set of all raters in the dataset.
- R_i: The subset of raters assigned to image i.

Disagreement, Distance and Cost

- d_{loc}: Localization distance (*e.g.*, $1 - \text{IoU}$).
- d_{cls}: Classification distance (0 for match, 1 for mismatch).
- d: A general distance function (used in KαLOS configuration).
- ψ: A general cost function for the correspondence solver.
- ψ_{soft}: A specific cost function that incorporates class information.
- C_{kl}: The cost of matching annotation k with annotation l.

KαLOS Pipeline Components:

- α or K-α: Krippendorff's Alpha.
- τ: Localization threshold.
- $\mathcal{S}$: The correspondence solver (*e.g.*, Greedy).
- M^*: The optimal correspondence set found by the solver.
- U_i: The set of disjoint "units" (clusters) for image i.
- NO_OBJECT: Special category for a rater's absence of an annotation in a unit. Active disagreement.

Noise Generator:

- λ: Magnitude of the noise generator.

8. Dataset Selection

The datasets considered in this work had to satisfy the following criteria:

- The task combines localization and classification.
- Annotations are fully human-created.
- The data are publicly accessible.
- A consistent annotation guideline is used across all annotations.

The core datasets that are eligible for at least parts of our study are summarized in Tab. 1. They are used either (1) for synthetic data creation (usable for empirical noise analysis), (2) as reference annotations for synthetic data generation and thus KαLOS validation, or (3) as examples for applications of KαLOS (see Sec. 14).

For completeness, we briefly discuss additional datasets that were considered but ultimately not used:

- **KeypointNet** [45]: designed for 3D keypoint detection, but only aggregated annotations are available, which makes it unsuitable for our multi-annotator analysis.
- **CorresPondenceNet** [26]: a 3D correspondence dataset

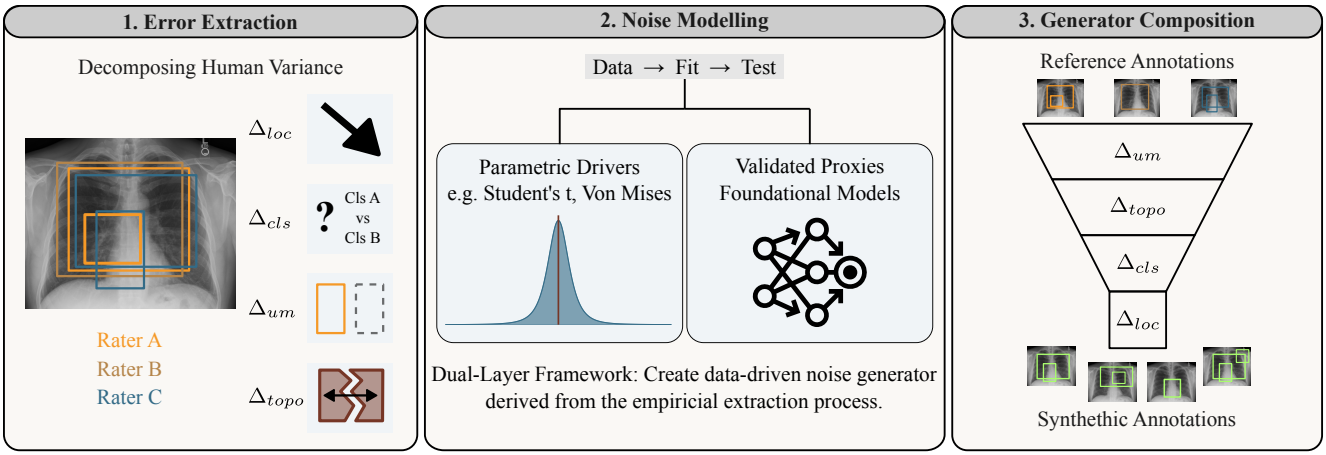

Figure 8. Three step noise generator: 1) Empirical data is extracted. 2) Data is used to fit parametric models or estimate them using foundational models as validated proxies. 3) The four separate components use a hierarchical composition to generate synthetic annotations based on reference annotations (*e.g.* human annotated data).

where annotators mark sets of semantically consistent points across ShapeNet objects, providing human correspondence consensus rather than explicit part or keypoint labels, which makes it unsuitable for our instance-based IAA setting.

- **IBIS** [42]: contains dental radiography landmarks, but the underlying data were no longer accessible at the time of our study.
- **MultiOrg** [6]: at the time of writing, the associated challenge was still ongoing, and the test images with repeated annotations were not yet publicly released.

9. Empirical Noise Generator

Goal and Scope. Validating agreement metrics requires an objective ground truth that mimics the complexity of human error. As Agnew *et al*. [1] note, the true distribution of annotation uncertainty remains unknown. Consequently, most existing methodologies rely on axiomatic heuristics or black-box machine generation, which fail to capture the heavy-tailed, non-isotropic nature of human disagreement. **The Dual-Layer Framework.** We address this gap by constructing a data-driven noise generator derived from the empirical extraction process of Tschirschwitz and Rodehorst [40]. We reject the assumption of uniform error. Instead, we propose a Dual-Layer Statistical Framework governed by a strict *Data* → *Fit* → *Test* loop:

1. **Parametric Drivers:** For geometric and topological errors, we fit empirical residuals to continuous distributions (*e.g.*, Student's t, Von Mises, Beta), validating the fit via Kolmogorov-Smirnov (KS) tests and the Akaike Information Criterion (AIC).
2. **Validated Proxies:** For semantic and visual ambiguity (Category Mistakes, False Positives), we utilize foundation models (CLIP [35], OWL-ViT [29]) as sampling distributions, validated against human confusion matrices using Mantel tests.

Data Sources. The generator is parameterized using pairwise comparisons extracted from the TexBiG [41], LVIS [16], and VinDr-CXR [32] datasets. While our experimental validation utilizes a generalized distribution averaged across these sources, the parameters can be estimated independently to create dataset-specific generators.

Section Structure. The remainder of this section details the construction of the generator:

1. We begin by reviewing prior noise generation models, highlighting the methodological leap from heuristics to empirical modeling (Sec. 9.1).
2. We describe the application of the error extraction pipeline [40], which isolates the raw disagreement data necessary for modeling (Sec. 9.2) see also Fig. 8.1.
3. We detail the core of our framework, systematically modeling each error type through our dual-layer approach (Sec. 9.3) as shown in Fig. 8.2.
4. Finally, we present the generator composition (Fig. 8.3), validating the pipeline by quantifying the signal loss (saturation) that occurs when high-priority topological errors consume candidates for lower-priority modifications (Sec. 9.4).

The full details of the generator can be found in the code: https://github.com/Madave94/empirical-vision-noise-generator

9.1. Review of Prior Noise Generation Models

Expanding upon Sec. 2.4, we provide a detailed comparison of previous synthetic noise generation models in Tab. 2. A general trend indicates that most methods for object detection remain relatively simplistic; notably, they often neglect

Table 2. Comparison of Synthetic Noise Generation Methods. We use the taxonomy of Mathet *et al*. [28] to model four channels: (i) **Loc.** (localization: shift, scale), (ii) **Unm.** (unmatched instances: missing, extra), (iii) **Cat.** (category mismatches), and (iv) **Fra./Com.** (fragmentation/combination). Without a reference, unmatched instances are sampled symmetrically. The study of Hu *et al*. [17] is excluded, since we didn't have any specifications of the generated noise. Note that Mathet *et al*. originally applied these concepts to text segmentation, not object detection, we present their idea for noise generation here as well since it's the most comprehensive.

Study	**Type**	**Modelling**	**Hypothesis (Assumptions)**
Agnew *et al*. [1]	Loc.	Uniform expansion (0-10 pixel buffering) of BBox	Noise is primarily expansion (outward shift).
	Loc.	Radial Gaussian noise $(1 - 5\mu)$ on mask vertices	Vertex precision varies normally; errors localize to edges.
Mathet *et al*. [28] (CST)	Unm.	Random dropout & addition (class is based on marginal dist.)	Errors are probabilistic; False Positives follow dataset statistics.
	Loc.	Random uniform shift	Shift magnitude is proportional to object length.
	Cat.	Confusion matrix (prevalence & overlap based)	Errors correlate with class frequency and semantic overlap.
	Fra.	Randomly select element and split it	Annotators mistakenly split single entities.
Bär *et al*. [7]	Loc.	Uniform noise $(\Delta h, \Delta w) \propto$ width/height	Error scales with object size; no directional bias. Same as [21]
Chachuła *et al*. [8] (CLOD)	Loc.	Random angle shift + scale change $\propto$ size	Errors involve position and size relative to dimensions.
	Unm.	Randomly adding spurious or deleting boxes	Presence/absence errors occur randomly.
	Cat.	Uniform label noise (random permutation)	Class confusion is random (no hierarchy assumed).
Li *et al*. [21]	Loc.	Uniform noise $(\Delta h, \Delta w) \propto$ width/height	Error scales with object size; no directional bias. Same as [7].
	Cat.	Symmetric noise (random flip)	Confusion probability is uniform across classes.
	Unm.	Proposals from a poorly trained detector (FP), this are machine generated label noise	FPs mimic realistic machine bias rather than random noise.
Liu *et al*. [24] (NLTE)	Cat.	Uniform random substitution	Class confusion is random.
	Unm.	Deletion if substituted label is "background" (FN)	Missing objects are misclassified as background.
Liu *et al*. [23]	Loc.	Uniform dist. (10-40%) relative to original bbox	Errors are strictly size-dependent relative shifts.
Chan *et al*. [10] (ODFI)	Unm.	Random removal; "Redundant" (duplicate + offset)	FPs are often "ghost" boxes (slight shifts of true objects).
	Loc.	Random reposition + size reduction (30%)	Errors involve deterministic shrinkage/displacement.
	Cat.	Superclass vs. Subclass swapping	Errors are hierarchically semantic (e.g. car↔truck $\neq$ car↔person).
Chadwick & Newman [9]	Cat.	Swap classes of nearby labels (Pair Noise) per image	Errors are systematic/spatially correlated (contextual confusion).
	Loc.	Shift and scale using Normal distribution	Spatial error is Gaussian; independent of object size.
	Unm.	Random addition (max 1/image) and removal	Spurious objects are rare events; missing is probabilistic.

complex error types such as combination and fragmentation, despite acknowledging their presence in real-world data [8]. The most comprehensive approach found in the literature is the Corpus Shuffling Tool (CST) proposed by Mathet *et al*. [28]. Although designed for textual data, its structural rigor serves as the foundation for our design.

Several assumptions are shared across multiple studies: (i) localization noise, specifically translation and scaling, is dependent on object size; (ii) errors follow uniform or normal distributions (*e.g*. class permutations, translation mag-

nitude/direction); and (iii) instance deletion (False Negatives) is modeled as an independent probabilistic event, unrelated to instance properties such as size or class. While individual heuristics may incorporate additional specific hypotheses, these three represent the core consensus across generation processes. A detailed breakdown is provided in Tab. 2.

9.2. Recapping Error Extraction

Tschirschwitz and Rodehorst [40] utilize the FiftyOne framework [30] to analyze annotation variation, identifying four error types that we map to Mathet's taxonomy [28]:

- Bounding Box Variation → Localization Shift
- Overlooked/Missed Instance → Unmatched Instance (FP/FN)
- Mismatched Class → Category Mistake
- Merged/Unmerged Instance → Fragmentation/Combination

The extraction pipeline operates on pairwise annotator matching, simplifying the model by ignoring higher-order interdependencies. Due to the lack of ground truth, errors such as unmatched instances are treated symmetrically; a discrepancy between annotators implies ambiguity regarding whether an instance is a false positive or false negative. The same happens for fragmentation and combination.

To account for matching sensitivity, we evaluate multiple IoU thresholds: 0.1, 0.25, 0.5, 0.75, and 0.9. As illustrated in Fig. 9, the "equilibrium point" where unmatched errors surpass conditionally matched errors varies significantly by dataset (*e.g.*, ≈ 0.75 for LVIS vs. ≈ 0.25 for VinDr-CXR). Rather than fitting a variable threshold per dataset, we maintain a fixed 0.5 threshold to ensure a standardized definition of "error" across our noise generator. This value aligns with the minimum matching criteria for mAP evaluation in both LVIS [16] and TexBiG [41]. While VinDr-CXR [32] utilizes a more lenient IoU of 0.4 for correct matches, we adhere to the 0.5 threshold to maintain parsimony and model consistency.

9.3. Noise Type-Specific Modeling

Note: To ensure scale invariance across images of varying resolutions, all geometric modeling is performed using relative coordinates normalized to the range $[0, 1]$ rather than absolute pixel values.

When we refer to reference annotations, we mean the annotations from which the generator derives the noisy annotation, usually a human annotated instance.

9.3.1. Localization Shift (Δ_{loc})

We classify localization shift as a *Parametric Driver*, a continuous error mode governing the geometric precision of the annotation. Unlike heuristic approaches that assume uniform noise, we model human spatial error as a superposition

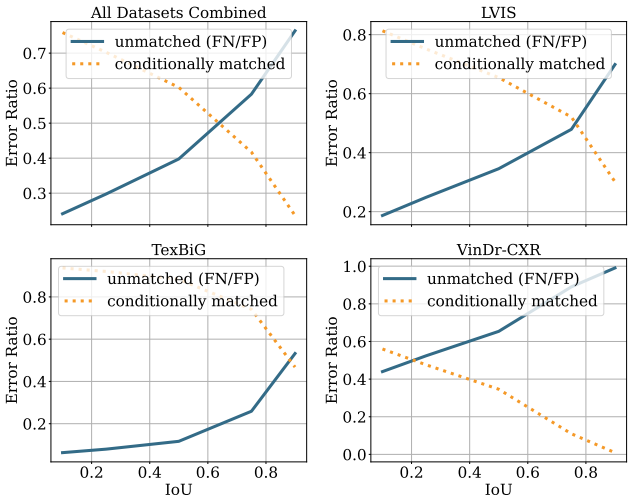

Figure 9. Impact of IoU threshold on error distribution. The plots illustrate the trade-off between unmatched instances (FP/FN) and conditionally matched instances (localization shifts, classification mistakes, combinations, and fragmentations) across the Combined, LVIS, TexBiG, and VinDr-CXR datasets.

of size-dependent trends and heavy-tailed stochastic residuals. We decompose this shift into three statistically independent components: translation magnitude, directional bias, and scale variation.

1. Translation Magnitude (Heavy-Tailed Residuals). Standard noise generators often assume translation error is normally distributed. Our empirical analysis rejects this assumption. While translation magnitude correlates linearly with object size, the residuals exhibit significant heavy tails ($KS_{Student-t} < KS_{Normal}$), indicating that extreme outliers occur more frequently than Gaussian models predict. Comparison of Log-Likelihood fits confirms that a Student's t-distribution provides a superior fit to the residuals. Consequently, we model the total translation magnitude Δ_{trans} as:

$$\Delta_{trans} = |(\alpha_t + \beta_t \cdot A_{avg}) + \text{Clip}(\mathcal{T}_\nu(\mu, \sigma), q_{0.001}, q_{0.999})| \tag{12}$$

where A_{avg} represents the average area of the paired instances, and α_t, β_t are coefficients derived from Multiple Linear Regression (MLR). $\mathcal{T}_\nu$ represents the random jitter drawn from the fitted Student's t-distribution. To prevent degenerate geometries during generation, we apply *Winsorizing*, strictly clipping the stochastic residuals to the $[0.001, 0.999]$ quantiles of the fitted distribution.

2. Directional Bias (The Cardinal Hypothesis). We scrutinize the common assumption that spatial error is isotropic (uniformly distributed in all directions). Hypothesis testing using both the Rayleigh test and the Kuiper test significantly rejects the null hypothesis of uniformity

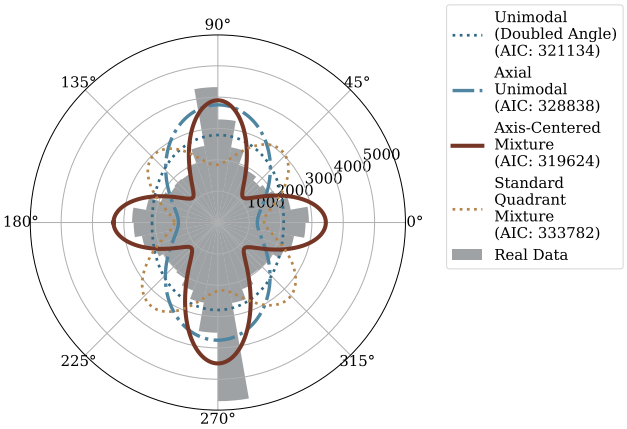

Figure 10. Analysis of directional bias in bounding box shifts. The gray histogram represents the empirical distribution of shift angles. We tested four parametric models: Unimodal, Axial Unimodal, Axis-Centered Mixture, and Standard Quadrant Mixture. The **Axis-Centered Mixture Model** (brown line) yields the lowest Akaike Information Criterion (AIC $\approx$ 319,623), validating the hypothesis that human error is biased toward cardinal axes ($0°, 90°, 180°, 270°$).

($p < 0.001$) across all datasets.

As illustrated in Fig. 10, human error exhibits a strong "Cardinal Bias," where annotators predominantly shift bounding boxes along the vertical and horizontal axes rather than diagonally. We model this probability density $P(\theta)$ using an Axis-Centered Mixture Model, defined as a weighted sum of four von Mises distributions centered at the cardinal directions:

$$P(\theta) = \sum_{k=1}^{4} w_k \cdot \text{VM}(\theta; \mu_k, \kappa_k) \tag{13}$$

where $\mu_k \in \{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}$ and the concentration parameters κ_k and weights w_k are estimated via Maximum Likelihood Estimation (MLE).

3. Scale Variation (Log-Space Asymmetry). Scale errors are multiplicative rather than additive; a 10-pixel error is negligible for a large object but catastrophic for a small one. We model width (s_w) and height (s_h) scaling as independent drivers in log-space. Similar to translation, we observe a size-dependent trend where smaller objects suffer larger relative scale errors (see Fig. 11).

We fit the absolute log-ratio of the reference and noisy dimensions to a linear model with Student's t residuals. For the width dimension, this is defined as:

$$\ln(s_w) = (\alpha_w + \beta_w \cdot A_{avg}) + \mathcal{T}_{\nu_w} \tag{14}$$

For general localization noise, the sign of the scaling factor is flipped with $p = 0.5$ to ensure symmetric expansion and contraction.

Engineering Constraints. To transition from theoretical distributions to a valid generator, we enforce two strict engineering constraints. First, as noted in Eq. (12), the heavy tails of the Student's t-distribution require Winsorizing to prevent physically impossible outlier generation. Second, we enforce a boundary clip: the centroid of any generated instance must remain within the image coordinates. If a stochastic update violates this condition, the vector is clamped to the image edge, ensuring valid topology for downstream tasks.

9.3.2. Category Mistake (Δ_{cls})

We model category mistakes as a hybrid error, integrating both the *Validated Proxy* and *Parametric Driver* strategies. While the decision of *which* class to substitute is driven by semantic proxies (visual embeddings), the physical realization of that error (the bounding box) is governed by re-calibrated parametric drivers.

1. The Semantic Transition Matrix (Validated Proxy). Standard noise generators often assume class confusion is uniform (random permutation). We reject this hy-

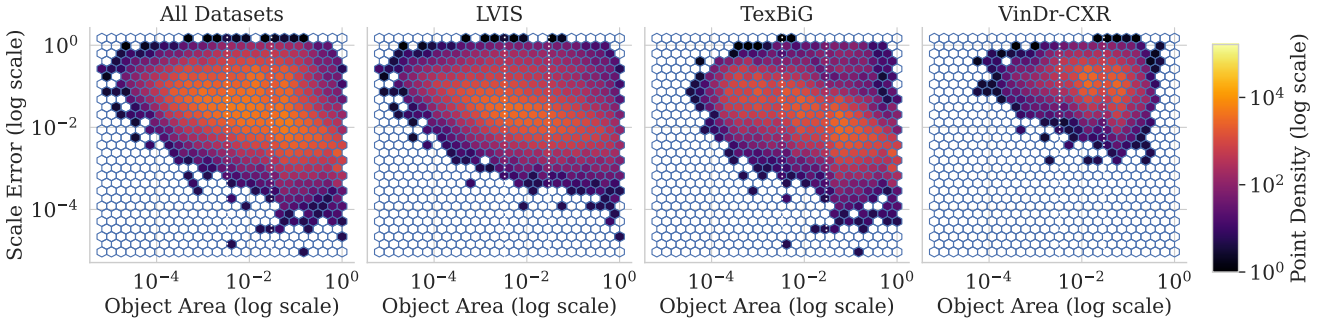

Figure 11. Relationship between relative scale error and object size. The hexbin plots visualize the density of scale errors across four datasets. A clear trend is visible where smaller objects (left side of x-axis) exhibit larger relative size differences (top of the y-axis), confirming the need for size-dependent noise modeling. The (thin) dashed white vertical lines divide into the COCO object size categories (adjusted to actual image size): Small ($< 32^2$), Medium (32^2 to 96^2), and Large ($> 96^2$) for an image of 640×480.

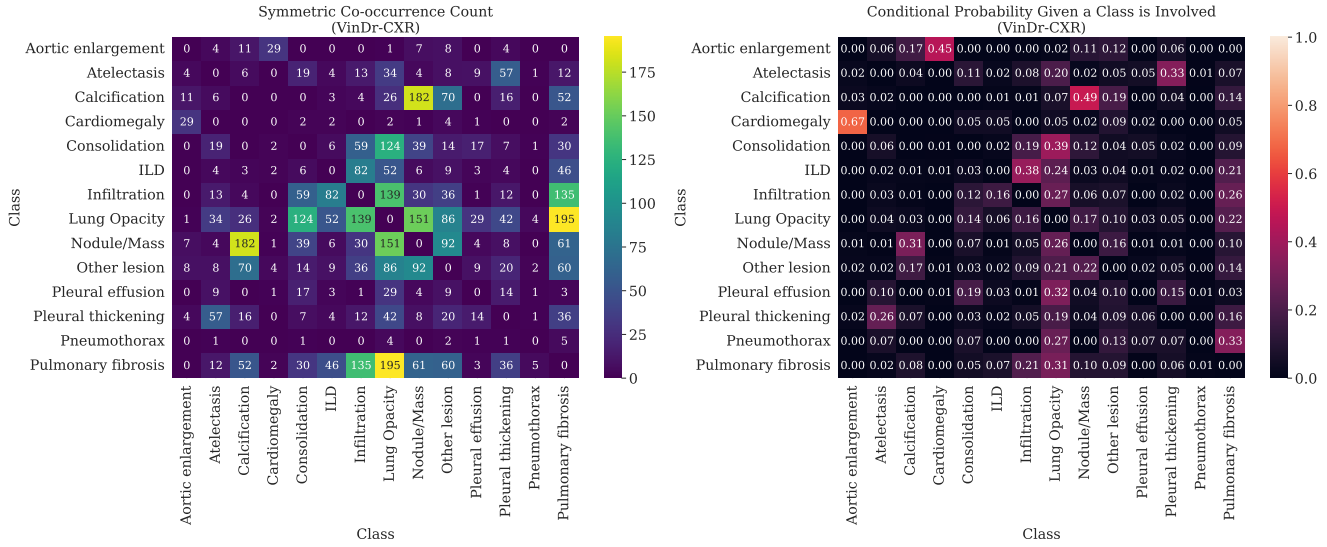

Figure 12. Analysis of classification co-occurrence for VinDr-CXR. Left: The symmetric (empirical) count of class confusions between annotators. Right: The conditional probability matrix, highlighting that confusion is often clustered among specific, semantically related pathologies (*e.g.*, Pulmonary fibrosis vs. Lung Opacity).

pothesis. A non-parametric Chi-squared permutation test on our data significantly rejects the null hypothesis of independence ($p < 0.05$). However, constructing robust empirical confusion matrices for long-tail datasets like LVIS is impossible due to data sparsity.

To address this, we utilize the visual embedding space of CLIP [35] as a validated proxy for human ambiguity. We compute a "Semantic Transition Matrix" based on the cosine similarity between the average visual centroids of all classes. We validate this proxy using a Mantel Test, which correlates the theoretical semantic similarity with the sparse empirical confusion matrix observed in the data. For LVIS, we observe a highly significant correlation ($r = -0.135, p = 0.0002$), confirming that semantic distance is a robust predictor of human labeling error.

In the generator, the probability of swapping the reference class c_{ref} with a new class c_{new} is modeled using a Softmax function over these similarities:

$$P(c_{new}|c_{ref}) = \frac{\exp(\text{sim}(\mathbf{v}_{c_{ref}}, \mathbf{v}_{c_{new}})/\tau)}{\sum_{c' \in \mathcal{C}_{top10}} \exp(\text{sim}(\mathbf{v}_{c_{ref}}, \mathbf{v}_{c'})/\tau)} \tag{15}$$

where $\mathbf{v}_c$ is the average CLIP embedding for class c, and $\text{sim}(\cdot)$ is the cosine similarity. We introduce a semantic temperature $\tau = 0.1$ to control the sharpness of the distribution, ensuring that confusions remain within a plausible semantic neighborhood (see Fig. 12).

2. Associated Localization Shift (Parametric Driver). Does a mislabeled object exhibit the same spatial error as a correctly labeled one? Our analysis suggests otherwise. We model the localization shift for misclassified instances using the same functional forms as Sec. 9.3.1 (Translation/Scale), but with re-estimated parameters.

Crucially, the directional bias changes. While standard localization error follows a 4-peak "Cardinal" distribution (snapping to x/y axes), misclassification errors are best fit by a Unimodal (Doubled Angle) von Mises distribution ($AIC \approx 11,752$) rather than the Axis-Centered Mixture ($AIC \approx 11,836$). This indicates that when annotators misidentify an object, their spatial precision degrades into a bi-modal axial distribution rather than the strict 4-way snapping observed in correct annotations.

3. Occurrence Rate. The decision to trigger a category mistake is modeled as a global Bernoulli trial, independent of object size, parameterized by the empirical pairwise error rate calculated across all datasets ($p_{global} \approx 0.026$).

9.3.3. Unmatched Instances (Δ_{um})

We model unmatched instances (errors of existence) as a hybrid error. While the decision of where to place a hallucinated object is driven by visual proxies (OWL-ViT [29]), the governing laws of how many and which sizes are affected are strictly parametric. Due to the lack of ground truth, we treat FP and FN symmetrically; a disagreement implies ambiguity regarding existence, not a definitive error by one party.

1. Occurrence Rate (Parametric Driver). We hypothesize that the frequency of unmatched instances correlates with scene complexity. To model this, we fit a Bayesian Poisson regression to the conditional count of unmatched instances given the total annotation density ($|Y_i^r|$) of the

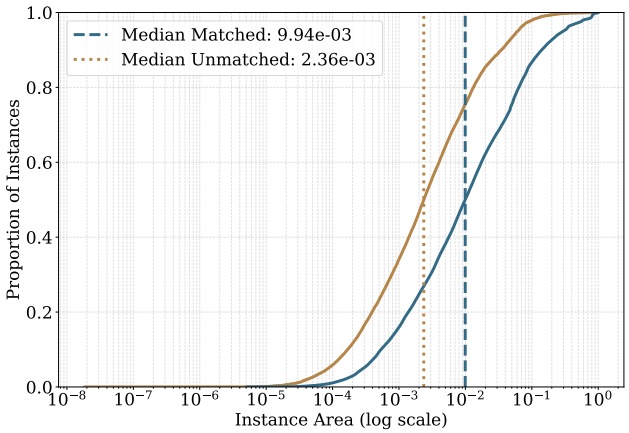

Figure 13. Cumulative Distribution Function (CDF) of instance areas for matched (blue) versus unmatched (orange) instances across all datasets. The distinct leftward shift of the unmatched curve indicates that missed or spurious annotations are systematically smaller than agreed-upon annotations.

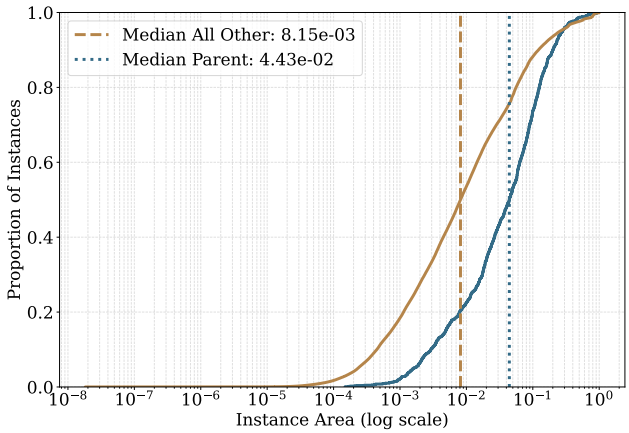

Figure 14. Cumulative Distribution Function (CDF) comparing the size of "parent" (blue) instances (involved in fragmentation/combination) versus all other instances. Parent instances are systematically larger, confirming that larger objects are more prone to split/merge errors.

rater. The expected number of noise events λ_{um} is modeled as:

$$\lambda_{um} = \exp(\alpha_{rate} + \beta_{rate} \cdot |Y_i^r|) \tag{16}$$

where α_{rate} and β_{rate} are the intercept and slope coefficients estimated via MCMC (Markov Chain Monte Carlo) sampling. Analysis of the posterior distributions ($\hat{R} \approx 1.0$) reveals distinct dataset characteristics: medical images (VinDr-CXR) show a steep error dependency ($\beta \approx 0.256$), while general scenes (LVIS) are more stable ($\beta \approx 0.021$). The final number of error events k is drawn from $Poisson(\lambda_{um})$.

2. Susceptibility and Pattern (Parametric Driver). Do annotators miss objects randomly? Our analysis rejects this heuristic. As shown in Fig. 13, unmatched instances are systematically smaller than matched instances. A Mann-Whitney U test confirms this size dependency is statistically significant ($p < 0.001$), proving that smaller objects are disproportionately prone to disagreement.

To capture this behavior, we fit a Logistic Regression on the log-transformed area of the instances. This yields a selection probability P_{select} for any given candidate (whether existing or proposed):

$$P_{select}(y) = \sigma(\alpha_{pat} + \beta_{pat} \cdot \ln(A_y)) \tag{17}$$

3. Visual Ghosts (Validated Proxy). For each of the k generated events, the system flips a coin ($p = 0.5$) to decide between deletion (FN) or addition (FP):

- **Deletion (FN):** Existing annotations are sampled for removal weighted by P_{select} (Eq. 17), ensuring that smaller objects are more likely to be "missed."
- **Addition (FP):** To generate realistic FP's, we cannot simply place random boxes. We utilize the open-vocabulary detector OWL-ViT [29] to generate a pool of visually plausible "ghost" proposals. We filter this pool to ensure validity: candidates must have an IoU < 0.1 with any existing annotation (to avoid creating duplicates). From this valid pool, a False Positive is sampled based on the size-weighted probability P_{select}. Finally, the class of the new instance is assigned using the Semantic Transition Matrix (CLIP [35] similarity) described in Sec. 9.3.2, ensuring the hallucinated object is semantically consistent with the scene context.

9.3.4. Fragmentation and Combination (Δ_{topo})

We classify topological errors (splitting one instance into two, or merging two into one) as a *Reversible Parametric Driver*. We utilize a single geometric model to govern both directions: it is used generatively to split parents into fragments and discriminatively to identify valid merge candidates via maximum likelihood.

1. Susceptibility (The Size Bias). Which objects are prone to topological disagreement? We hypothesize that complexity scales with size. A Mann-Whitney U test confirms that "parent" instances (those involved in split/merge errors) are statistically significantly larger than the general population ($p < 0.001$, see Fig. 14). We model this susceptibility P_{parent} using a Logistic Regression on the log-transformed area:

$$P_{parent}(y) = \sigma(\alpha_{frag} + \beta_{frag} \cdot \ln(A_y)) \tag{18}$$

2. The Geometry of Rupture (Generative). Once a parent is selected for fragmentation, we generate two children using a two-step dependency chain.

(a) **Parent-to-Child (Asymmetric):** The first child is generated using a modified localization model. Unlike standard shifts, scaling parameters are conditioned on the *parent's* area (not average area) and are strictly negative (reduction).
Calibration Note: Validation against empirical data revealed that Maximum Likelihood Estimation systematically underestimated child sizes (Median area ratio: 0.059 synthetic vs. 0.088 real). To correct this, we apply a post-hoc empirical calibration: a scalar factor ($\kappa \approx 0.20$) is added to the log-scaling intercepts, derived from the square root of the median discrepancy, ensuring synthetic fragments sum to a realistic proportion of the parent.

(b) **Child-to-Child (The "Opposing Sides" Hypothesis):** The second child is positioned relative to the first. We hypothesize that fragments typically represent opposing parts of an object (*e.g.* head vs. torso). Empirical analysis confirms that 87.3% of fragment pairs reside on opposing sides of the parent's major axis. We model the angular difference θ_{cc} between the two child vectors using a Beta Distribution:

$$\theta_{cc} \sim \text{Beta}(\alpha_\theta = 4.53, \beta_\theta = 0.53) \tag{19}$$

As shown in Fig. 15, the distribution peaks near $180°$ (normalized to 1.0), strictly enforcing the spatial separation observed in real disagreement.

3. Combination via Model Inversion (Discriminative). To simulate combination (merging two instances), we invert the generative logic. Rather than using heuristics, we calculate a Geometric Likelihood Score $S(y_a, y_b)$ for every pair of same-class annotations. This score represents the joint probability density that the two instances *could have been generated* as fragments of the same parent:

$$S(y_a, y_b) = f_t(\Delta_{trans}) \cdot f_w(s_w) \cdot f_h(s_h) \cdot f_\theta(\theta_{rel}) \tag{20}$$

where $f(\cdot)$ are the Probability Density Functions (PDF)

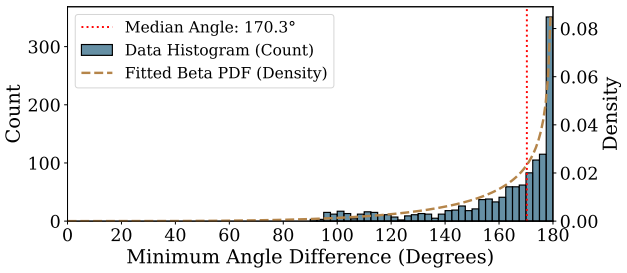

Figure 15. Distribution of the angular difference between two child fragments relative to their parent centroid. The peak near $180°$ indicates that fragments tend to be located on opposing sides of the object, rather than clustered together.

of the fitted Student's t (translation/scale) and Beta (angle) distributions. Pairs with high likelihood scores are merged into a single bounding box, ensuring that synthetic merges respect the statistical properties of human annotation patterns.

9.4. Generator Composition

We combine the isolated parametric drivers and validated proxies into a unified **Empirical Noise Generator**. To resolve logical conflicts, such as the impossibility of shifting a bounding box that has been deleted, we enforce a strict topological hierarchy. The generator operates sequentially on the set of clean annotations Y, prioritizing errors that alter existence and topology over those that alter attributes or precision.

1. The Execution Hierarchy. The generation pipeline executes four mutually exclusive stages. We maintain a dynamic registry of valid candidates; as high-priority processes consume "clean" annotations, the pool available for subsequent steps diminishes.

1. **Unmatched Instances (Priority 1: Existence).** The generator first determines the set of active instances. FN remove candidates from the pool, while FP inject new proposals directly into the completed set, bypassing subsequent modification stages to preserve their "ghost" status.
2. **Fragmentation/Combination (Priority 2: Topology).** Remaining candidates are evaluated for split/merge events. Successful topological changes consume the original parents and produce new, geometrically distinct children.
3. **Category Mistake (Priority 3: Semantics).** The remaining clean instances are sampled for class permutation. Crucially, as noted in Sec. 9.3.2, misclassified instances receive their own specific localization noise profile and are effectively removed from the standard localization pool.
4. **Localization Shift (Priority 4: Geometry).** Finally, any annotation that has survived the preceding stages without modification is subjected to the baseline localization shift (translation and scale) defined in Sec. 9.3.1.

2. Magnitude Control (λ). The global noise intensity is controlled by a scalar factor λ. This parameter scales both the frequency of discrete error events and the variance of continuous geometric errors.

- **Rate Scaling:** For discrete events (*e.g.* category mistake), the empirical probability p_{base} is scaled such that $p_{final} = \min(1.0, p_{base} \cdot \lambda)$.
- **Geometric Scaling:** For continuous drivers, λ acts as a multiplicative factor on the stochastic residuals drawn from the Student's t distributions: $\Delta_{final} = \Delta_{fitted} \cdot \lambda$.

3. Signal Loss (Cannibalization). A consequence of this hierarchy is *Signal Loss*. As λ increases, high-priority

Figure 16. Visual comparison of synthetic (solid lines) and reference (dashed lines) annotations in LVIS. Matched pairs are shown in light orange (synthetic) and green (reference). Unmatched synthetic annotations (false positives) are navy, while unmatched reference annotations (false negatives) are brown. Images are cropped to a quadratic format.

processes (like deletion or fragmentation) consume a larger proportion of the available candidates. We track a metric termed "cannibalization", which counts the number of times a lower-priority process attempts to select a candidate that no longer exists.

As shown in Tab. 3, at $\lambda = 1.0$ (empirical baseline), signal loss is minimal (4.2%). However, at extreme magnitudes ($\lambda = 5.0$), nearly 30% of theoretically sampled errors are "cannibalized." This saturation is not a flaw but a valid simulation of information entropy: as noise overwhelms the signal, the distinctions between specific error types vanish, leaving only a chaotic distribution.

Table 3. Signal Loss Analysis (LVIS). As noise magnitude λ increases, the competition for finite candidates leads to saturation, where the effective error rate diverges from the theoretical sampling rate.

Magnitude λ	0.25	0.50	1.00	2.00	5.00
Signal-loss ratio	0.2%	1.3%	4.2%	10.1%	28.3%

Limitations. We identify three primary limitations of this framework. (1) *Saturation at high magnitudes:* As λ increases, the competition for unmodified candidates intensifies, leading to signal loss. At extreme magnitudes ($\lambda \gg 2.0$), the reference signal degrades until the output resembles random noise. This saturation is expected behavior, as noise levels significantly beyond the empirical baseline eventually obscure the underlying signal entirely. (2) *Visual independence:* With the exception of proposal based false positive generation, the stochastic processes are decoupled from image content. While factors such as occlusion or blur drive real world human error, explicit modeling of these visual cues was not required for the validation of agreement metrics. Future iterations targeting model training data augmentation would benefit from incorporating such uncertainty maps. (3) *Validation scope:* Our validation focuses on the goodness of fit between empirical data and our parametric models. While a direct statistical comparison between the generated synthetic distributions and the original empirical distributions remains a subject for future investigation, this framework nevertheless significantly advances the state of the art by modeling complex and interdependent error modalities rather than relying on isolated heuristics or black-box machine label noise.

10. Comparing Cost Functions

To compare the effect of incorporating semantic information into the correspondence cost, we evaluate a baseline cost function that relies solely on localization agreement. This cost, denoted ψ_{neg}, assigns higher preference to pairs with greater spatial overlap but remains indifferent to class identity. Formally, it is defined as:

$$\psi_{\text{neg}}(d_{\text{loc}}, d_{\text{cls}}) = -d_{\text{loc}}, \tag{21}$$

The class aware cost function is in the main paper as Eq. (6).

The qualitative example in Fig. 17 illustrates a broader phenomenon frequently encountered in instance-based vision tasks: multiple semantic categories may legitimately occupy nearly identical spatial regions. When localization alone is used as the correspondence signal, such situations induce ambiguity; spatial overlap becomes insufficient to determine which annotations should be matched. In these cases, solvers that ignore class information are prone to semantically implausible correspondences, especially under realistic annotation noise.

Figure 17. Rater-level annotations for LVIS image No. 467776 (shows only the region of interest) illustrating a conceptual edge case where class information must influence the correspondence cost. The two panels on the left show annotations from rater r1, and the two on the right from rater r2. Both raters provide two labels (cow and calf) for the very same physical instance, and both classes occupy effectively identical spatial regions. Because object detection allows overlapping instances, the spatial component alone provides no discriminative signal: all four boxes overlap nearly perfectly, and localization-based matching is therefore ambiguous. A class-aware cost function resolves this ambiguity in a principled way by preferring correspondences that preserve each rater's intended class assignment rather than artificially swapping them. This example illustrates why incorporating class information into the matching cost is conceptually necessary for certain real-world annotation patterns, even before considering any quantitative evaluation.

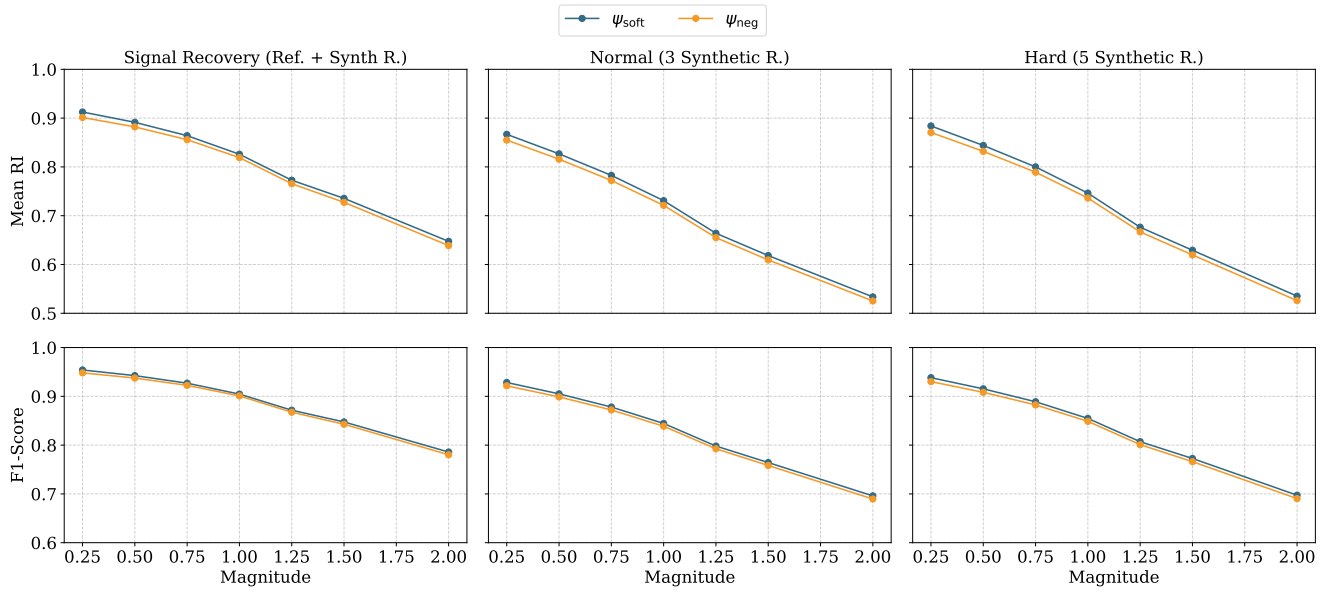

Figure 18. Comparison of the two correspondence cost functions on LVIS using the Greedy solver across noise magnitudes and rater configurations. The category-lenient cost ψ_{soft} consistently achieves slightly higher filtered Rand Index and F1-scores than the baseline ψ_{neg}. While the differences are small, they are systematic and stable across all settings, indicating that incorporating class information into the cost improves correspondence robustness.

Figure 18 quantitatively confirms this intuition. Across all noise magnitudes and rater configurations, the category-lenient cost ψ_{soft} consistently outperforms the purely localization-based ψ_{neg}. While the numerical differences are modest, they are systematic and robust: ψ_{soft} reliably avoids precision-degrading errors in which spatially similar but semantically incompatible annotations are merged. These findings show that even a simple class-consistency term stabilizes correspondence formation, particularly when spatial overlap is high due to class definitions that allow or encourage closely colocated instances.

11. Robustness Analysis of Data-Driven Calibration

To evaluate the stability of the distance metric d and calibration anchor τ^*, we perform a non-parametric percentile bootstrap. We resample N images with replacement over 100 iterations to compute 95% confidence intervals. Be-

Table 4. Global robustness analysis of the calibration anchor (τ^*) and KS statistic across 100 bootstrap iterations (All Sizes). CI denotes the 95% Confidence Interval.

Dataset	Annotations	τ^* Mean (95% CI)	KS Mean (95% CI)
LVIS [16]	100,480	0.654 ([0.643, 0.668])	0.798 ([0.786, 0.807])
TexBiG [41]	52,372	0.435 ([0.420, 0.455])	0.803 ([0.796, 0.810])
VinDr-CXR [32]	36,096	0.996 ([0.996, 0.997])	0.684 ([0.675, 0.692])

cause calibration is dataset-dependent, we apply this analysis to object detection on VinDr-CXR [32], TexBiG [41], and LVIS [16] using IoU. As Tab. 4 demonstrates, τ^* remains highly stable across samples. This tight variance confirms that the empirical noise profile is a consistent property of the dataset, justifying minor threshold adjustments (e.g., rounding to align with standard IoU benchmarks) provided they fall near these confidence bounds.

Standard IoU correlates poorly with human perception, disproportionately penalizing localization errors in smaller objects [25, 39]. To investigate this bias, we stratify our robustness analysis by relative instance size (small, medium, and large, scaled proportionally from standard COCO definitions to relative image area). We employ an asymmetric matching strategy: target instances of a specific size class are matched against the entire pool of reference annotations to prevent artificial boundary exclusion.

Tab. 5 reveals a pronounced scale dependency across all domains. Smaller objects consistently require significantly more lenient (higher) distance thresholds to successfully separate signal from noise, whereas larger objects demand much stricter tolerances. Consequently, selecting a single global similarity threshold imposes a fundamental trade-off: stricter thresholds accurately capture geometric precision but discard valid small-object annotations as noise, while lenient thresholds capture small objects but misclassify stochastic chance overlaps as true consensus for large objects.

Crucially, this scale bias is an inherent flaw of the chosen distance metric (IoU), not KαLOS. Because KαLOS operates as a modular meta-algorithm, it successfully isolates and exposes this metric bias. Future applications can mitigate this issue entirely by substituting IoU with a scale-invariant distance metric within the KαLOS framework.

12. Global vs Mean K-α

The primary KαLOS metric is computed on a per-image basis, and the dataset-wide score is reported as the mean of these values (mean K-α). We empirically validate this image-centric approach because it ensures that agreement is balanced across independent scenes.

An alternative approach is to concatenate all per-image reliability matrices and run a single calculation on the resulting global matrix. We refer to this unvalidated variant as the secondary global metric (global K-α). While global K-α can be useful for datasets with highly uniform instance counts per image or for cases where agreement on absence is intentionally weighted lower, we treat it strictly as a secondary diagnostic due to inherent statistical biases:

- **Instance-Based Domination:** global K-α weights the evaluation by instance count rather than by scene. A single dense image containing 100 instances will mathematically overwhelm the consensus of 100 sparse images containing one instance each. This disproportionately biases the metric toward the geometric precision of clustered ob-

Table 5. Size-stratified robustness analysis using asymmetric matching. Small objects require highly lenient (higher) distance thresholds compared to large objects, exposing the scale bias of IoU.

Dataset	Size	Target Anns	τ^* Mean (95% CI)	KS Mean (95% CI)
LVIS [16]	Small	35,574	0.976 ([0.973, 0.978])	0.773 ([0.750, 0.795])
	Medium	38,558	0.748 ([0.725, 0.771])	0.863 ([0.855, 0.870])
	Large	26,348	0.488 ([0.473, 0.501])	0.889 ([0.883, 0.894])
TexBiG [41]	Small	12,769	0.695 ([0.644, 0.764])	0.904 ([0.894, 0.918])
	Medium	15,376	0.607 ([0.570, 0.651])	0.880 ([0.870, 0.889])
	Large	24,227	0.117 ([0.113, 0.121])	0.843 ([0.834, 0.851])
VinDr-CXR [32]	Small	3,976	0.995 ([0.995, 0.996])	0.524 ([0.497, 0.553])
	Medium	17,262	0.996 ([0.996, 0.997])	0.659 ([0.649, 0.667])
	Large	14,858	0.875 ([0.848, 0.956])	0.776 ([0.768, 0.785])

jects (e.g., dense document layouts) while masking disagreement in sparse scenes.

- **Distortion of Expected Disagreement (D_e):** Concatenating matrices merges the class distributions of all independent scenes. The metric's baseline for chance agreement (D_e) becomes skewed by the most frequently annotated classes in the densest images, artificially altering the penalty for misclassification across the rest of the dataset.
- **Devaluation of Absence:** In domains like medical analysis, an entirely empty image represents a critical consensus on the absence of a finding. In mean K-α, this perfect agreement holds equal weight to an annotated scan. In global K-α, to mitigate the issue of the matrix dropping the empty image entirely, it is injected as a single virtual `NO_OBJECT` agreement. While this reduces the issue it still remains. Consequently, an image with 50 annotations carries 50 times the mathematical weight of a clinically significant empty image.

13. Downstream Tasks

Regardless of the task of computer vision, KαLOS arrives at a nominal reliability matrix. Since the data structure is always the same, downstream diagnostics (Fig. 2.4) become available. By re-computing α on filtered or permuted views of this matrix, the framework enables granular diagnostics beyond a single summary score.

Per-Image Agreement Distribution: Plotting the per-image α scores (see Fig. 1c) reveals the stability of the consensus. A left-skewed distribution identifies specific problematic images or ambiguous scenarios, enabling targeted refinement of annotation guidelines during dataset creation.

Localization Sensitivity Analysis: This analysis quantifies the agreement "lost" to boundary imprecision. By computing the delta between agreement at the calibration anchor (τ^*) and a stricter threshold (τ_{strict}), *i.e.*, $\Delta = \alpha(\tau^*) - \alpha(\tau_{strict})$, we isolate true geometric jitter from fundamental existence disagreement.

Class Recognition Difficulty: We diagnose semantic ambiguity by filtering the reliability matrix for a single class c. Crucially, under the completeness assumption, any rater who failed to annotate c is assigned NO_OBJECT. Re-computing α on this subset distinguishes well-defined classes (*e.g.*, `person`: $\alpha \approx 0.8$) from ambiguous ones (*e.g.*, `distant bird`: $\alpha \approx 0.4$).

Collaboration Clusters: This analysis identifies "schools of thought" by generating a pairwise vitality matrix. High pairwise agreement between subsets of raters reveals implicit conventions or institutional biases distinct from the global consensus.

Intra-Annotator Agreement: By treating a rater's annotations at times t_0 and t_1 as independent rows in the reliability matrix, KαLOS quantifies self-consistency without architectural changes.

Annotator Vitality: Following Nassar *et al*. [31], we measure an individual rater's contribution to the consensus. The vitality V_r for rater r is defined as:

$$V_r = \alpha_R - \alpha_{R\setminus\{r\}} \tag{22}$$

where α_R is the score with the full cohort and $\alpha_{R\setminus\{r\}}$ is the score with rater r removed. A positive V_r identifies a consensus builder; a negative V_r identifies a source of noise or deviation.

14. Application of KαLOS

This section demonstrates the adaptability of the KαLOS meta-algorithm across distinct computer vision domains. For each task, we derive the principled configuration via distributional analysis, identifying the optimal distance function (d_{loc}) and the calibration anchor (τ^*) required to statistically separate annotator signal (D_o) from stochastic noise (D_e).

Crucially, we apply the completeness assumption globally for these experiments: we assume that raters annotate all instances they perceive. Consequently, if a rater fails to annotate a unit discovered by the group, it is encoded as an explicit active disagreement (NO_OBJECT), penalizing the score as a false negative, rather than treated as missing data. We present this diagnostic pipeline on three diverse tasks: (1) instance segmentation on TexBiG [41], (2) 3D volumetric segmentation on LIDC-IDRI [3], and (3) pose estimation on MARS [38].

Note: For the distributional analysis distances d are used. However, it is common practice in most tasks (object detection, instance segmentation) to use similarities instead. Hence, for the distributional analysis, we provide distance (and as a secondary axis similarity values), but in later analysis similarities are used in accordance with standard practice. The distance threshold is shown as τ and similarity thresholds as τ_s.

14.1. Instance Segmentation

TexBiG [41] is a dataset on complex layouts of historical documents with dense annotations covering all elements in the layout beside the background (mostly white).

Principled Configuration. To adapt KαLOS for instance segmentation, we first determine the optimal localization distance function d_{loc} and the associated calibration anchor τ^*. We evaluate three candidate functions: Polygon IoU, Mask GIoU, and L2 Centroid distance. As the framework requires a distance metric d, we normalize and invert similarity scores where necessary. Annotations are defined as $\tilde{y}^r_{ik} = (\tilde{p}^r_{ik}, \tilde{c}^r_{ik})$, where $\tilde{p}^r_{ik}$ represents the segmentation polygon and $\tilde{c}^r_{ik}$ remains the class, same as for bounding boxes.

1. **Polygon IoU:** We calculate the standard intersection over union for polygonal masks. To convert this similarity into a distance, we invert it:

$$d_{\text{IoU}}(\tilde{p}^a_{ik}, \tilde{p}^b_{il}) = 1 - \frac{|\tilde{p}^a_{ik} \cap \tilde{p}^b_{il}|}{|\tilde{p}^a_{ik} \cup \tilde{p}^b_{il}|} \tag{23}$$

2. **Mask GIoU:** The Generalized IoU (GIoU) accounts for spatial proximity in non-overlapping shapes.

$$\text{GIoU}(\tilde{p}^a_{ik}, \tilde{p}^b_{il}) = \text{IoU}(\tilde{p}^a_{ik}, \tilde{p}^b_{il}) - \frac{|C \setminus (\tilde{p}^a_{ik} \cup \tilde{p}^b_{il})|}{|C|}, \tag{24}$$

where C is the smallest convex hull enclosing both $\tilde{p}^a_{ik}$ and $\tilde{p}^b_{il}$. Since the standard GIoU range is $[-1, 1]$, we first normalize it to a similarity $S \in [1]$ and then invert it to define the distance:

$$d_{\text{GIoU}} = 1 - \frac{1 + \text{GIoU}}{2} \tag{25}$$

3. **L2 Centroid Distance:** We compute the Euclidean distance between geometric centroids defined as $\tilde{\mathbf{p}}^{\mathbf{r}}_{\mathbf{ik}}$, normalized by the image diagonal to ensure the range $[0, 1]$.

$$d_{L2}(\tilde{p}^a_{ik}, \tilde{p}^b_{il}) = \|\mathbf{m}(\tilde{p}^a_{ik}) - \mathbf{m}(\tilde{p}^b_{il})\|_2. \tag{26}$$

We identify the optimal metric to measure the annotator intention by maximizing the statistical separation between observed disagreement (D_o) and expected chance disagreement (D_e) as shown for IoU in Fig. 19. We quantify this separation using the Kolmogorov-Smirnov (KS) statistic. Tab. 6 presents the results for the TexBiG dataset.

Distance Metric	KS Statistic ↑	Calibration Anchor τ^*
Polygon IoU	**0.82**	0.53
L2 Centroid	0.78	0.02
Mask GIoU	0.78	0.28

Table 6. TexBiG results for principled configuration.

The data indicates that Polygon IoU achieves the highest separation (KS=0.8189), confirming it as the most robust metric for distinguishing annotator consensus from random chance in this domain. While the statistically optimal calibration anchor is $\tau^* = 0.532$, we adopt the standard threshold $\tau = 0.5$. We select this value because it lies within the statistical margin of error (See robustness in Sec. 11) of τ^* and maintains alignment with standard computer vision benchmarks (*e.g.*, mAP@50), facilitating cognitive comparison for future research.

Specification and dataset-wide mean Agreement. Based on this calibration, we define the complete configuration as:

$$\text{K}\alpha\text{LOS}\big(d_{\text{IoU}}, \tau{=}0.5, \mathcal{S}{=}\text{Greedy}, \psi_{\text{soft}}\big) \tag{27}$$

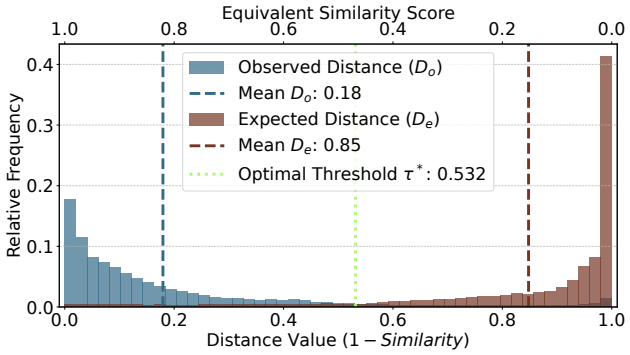

Figure 19. Distribution of Observed (D_o) vs. Expected (D_e) disagreement for instance segmentation on TexBiG using $1 - \text{IoU}$.

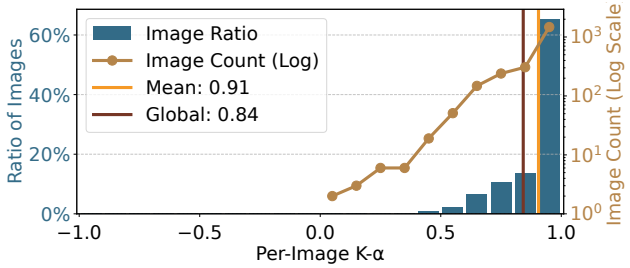

Figure 20. **Per-Image Agreement.** The left-skewed distribution reveals that while the median image achieves perfect consensus ($\alpha = 1.0$), a long tail of disagreement exists, identifying specific ambiguous samples for review.

Executing this configuration on the TexBiG dataset yields a dataset-wide mean K-α of **0.9055**. This score indicates "almost perfect" agreement, but a scalar summary obscures the specific sources of disagreement. By standardizing the complex instance segmentation task into a nominal reliability matrix, the framework unlocks granular diagnostics regarding data ambiguity and annotator behavior.

Data Diagnostics. The distribution analysis in Fig. 20 confirms that the consensus is stable, with a median $\alpha = 1.0$. However, the Localization Sensitivity Analysis (LSA) in Fig. 21 exposes the physical limits of this consensus. Agreement remains high (> 0.90) for thresholds $\tau_s \in [0.1, 0.5]$, indicating raters agree on object existence. The sharp drop at $\tau_s = 0.9$ ($\alpha = 0.35$) proves that pixel-perfect boundary consensus is unachievable in this domain. Furthermore, Fig. 22 isolates semantic ambiguity. While structural classes like `header` ($\alpha = 0.99$) are solved, abstract concepts like `equation` ($\alpha = 0.45$) generate significant noise, pinpointing exactly where the annotation guidelines fail.

Annotator Diagnostics. The framework isolates human factors without requiring ground truth. Annotator Vitality (Fig. 23) identifies `coder_d` as a stabilizer who increases global agreement by 0.04, whereas `coder_e` is

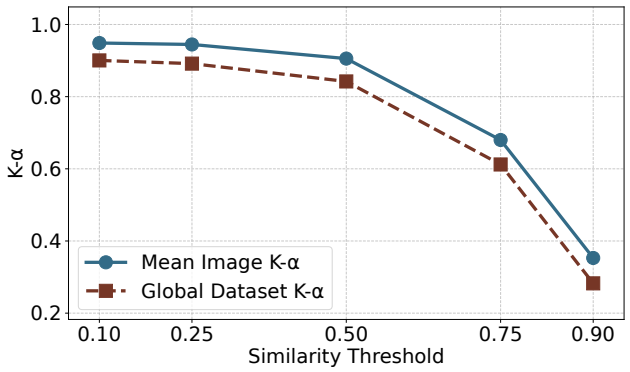

Figure 21. **Localization Sensitivity Analysis (LSA).** Agreement remains stable ($\alpha \approx 0.94 \rightarrow 0.90$) up to $\tau_s = 0.5$ but collapses at $\tau_s = 0.9$ ($\alpha = 0.35$). This "Cliff of Precision" quantifies the limit of human spatial consistency.

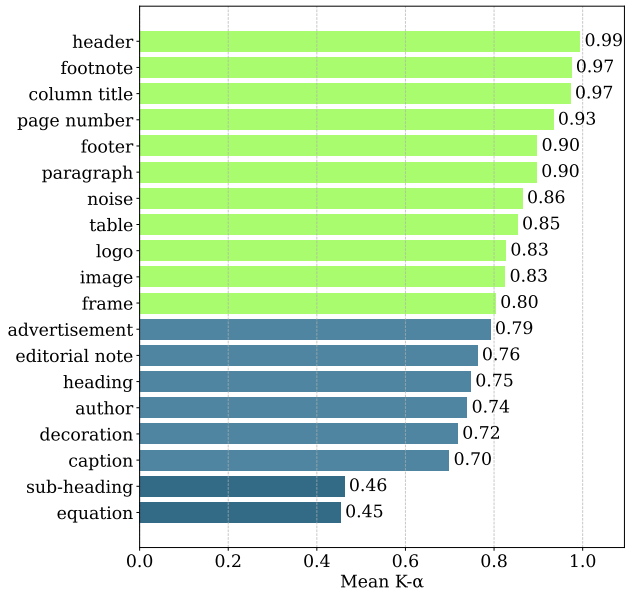

Figure 22. **Class Recognition Difficulty.** Structural elements like `header` ($\alpha = 0.99$) are robust, whereas semantic definitions for `equation` ($\alpha = 0.45$) and `sub-heading` ($\alpha = 0.46$) require guideline refinement.

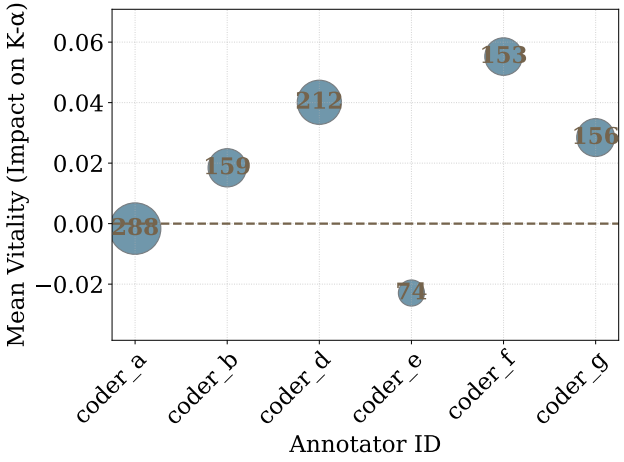

Figure 23. **Annotator Vitality.** `coder_d` ($V_r = +0.04$) acts as a strong consensus builder, while `coder_e` ($V_r = -0.02$) introduces systematic noise, suggesting a need for retraining.

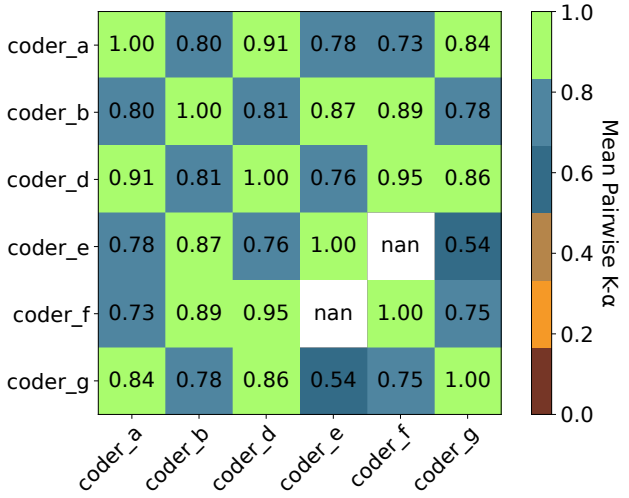

Figure 24. **Collaboration Clusters.** The heatmap reveals distinct "schools of thought." `coder_f` and `coder_d` achieve higher pairwise agreement (0.95) than the global average, indicating shared implicit conventions.

a source of divergence ($V_r = -0.02$). Finally, the Collaboration Heatmap (Fig. 24) detects implicit "schools of thought." The high pairwise agreement between `coder_f` and `coder_d` (0.95) versus `coder_e` (0.76) suggests divergent interpretations of the task, allowing for targeted intervention rather than dataset pruning.

14.2. 3D Volumetric Instance Segmentation

To demonstrate the transferability of KαLOS from 2D images to 3D volumetric data, we apply the meta-algorithm to the LIDC-IDRI dataset [3]. This task involves the binary classification (Nodule vs. Background) and localization of pulmonary nodules in CT scans. We restrict the analysis to nodules ≥ 3 mm, as smaller instances lack explicit contour definitions in the source data. Furthermore, because LIDC-IDRI anonymizes annotator identities, we treat the four reading sessions per scan as independent, unidentifiable raters.

Principled Configuration. We adapt the distance metric to operate on voxel grids rather than pixel masks. Let V_{ik}^r represent the set of voxels occupied by annotation k from rater r in volume i. We define the Volumetric IoU distance

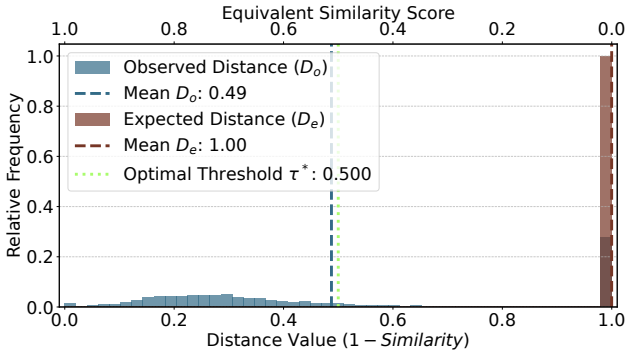

Figure 25. **Signal vs. Chance in 3D.** The distribution of Expected Disagreement (D_e) converges to a Dirac delta at 1.0. This reflects the sparsity of nodules within the vast 3D volume; unlike in 2D images, the probability of two random annotations overlapping in a voxel grid is statistically negligible.

d_{vol} as:

$$d_{\text{vol}}(V_{ik}^a, V_{il}^b) = 1 - \frac{|V_{ik}^a \cap V_{il}^b|}{|V_{ik}^a \cup V_{il}^b|}. \tag{28}$$

To determine the validity of this metric, we analyze the statistical separation between the Observed Disagreement (D_o) and Expected Disagreement (D_e). As shown in Fig. 25, the Kolmogorov-Smirnov (KS) statistic maximizes at a threshold of $\tau^* = 0.50$ (KS $= 0.7237$). While one could search for alternative distance functions to further maximize this separation, the high KS score confirms that d_{vol} successfully distinguishes genuine inter-rater correspondence from chance. Consequently, we settle on this metric and update the calibration anchor to the data-driven optimum:

$$\text{K}\alpha\text{LOS}_{\left(d_{\text{vol}},\, \tau=0.50,\, \mathcal{S}=\text{Greedy},\, \psi_{\text{soft}}\right)}. \tag{29}$$

Domain-Specific Topology. The distribution analysis of D_o vs D_e in Fig. 25 reveals a unique characteristic of volumetric analysis: the Expected Disagreement (D_e) converges strongly to 1.0. In 2D tasks, random boxes often overlap slightly, creating a "soft" chance distribution. In 3D, the minute spatial footprint of a nodule relative to the CT volume makes random overlap impossible. This simplifies the problem topology: since chance is effectively constant ($D_e \approx 1$), the reliability of the dataset depends entirely on the precision of the Observed Disagreement (D_o).

Results and Interpretation. The pipeline yields a dataset-wide mean agreement of $\alpha = 0.3683$. While significantly lower than the "Almost Perfect" scores observed in TexBiG (0.9055), this does not imply poor data quality. Rather, it quantifies the inherent ambiguity of the medical domain. As seen in the Per-Image Distribution (Fig. 26), agreement varies wildly, with a heavy tail of difficult cases

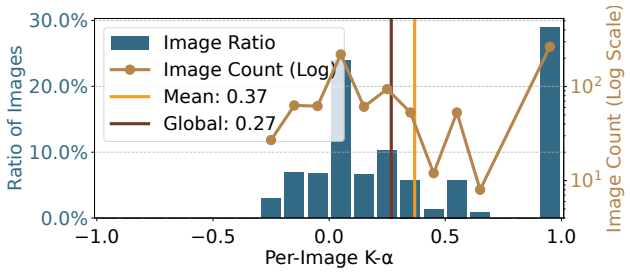

Figure 26. **Per-Image Agreement Distribution.** The distribution shows a significant spread (Mean $\alpha = 0.37$). While a subset of scans achieves perfect consensus ($\alpha = 1.0$), the heavy tail indicates that medical annotation involves inherent ambiguity not present in structural tasks like document layout.

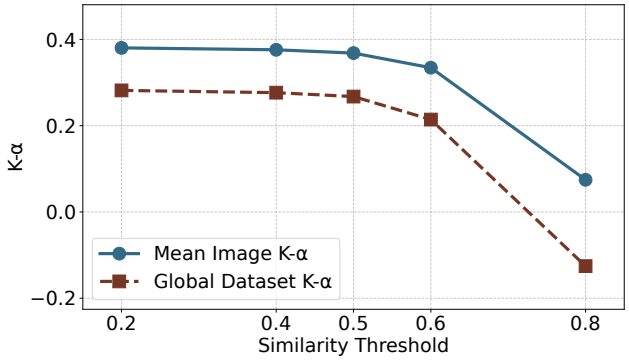

Figure 27. **Localization Sensitivity (LSA).** The agreement exhibits a "Plateau" behavior, remaining stable ($\alpha \approx 0.37$) up to $\tau_s = 0.50$. Agreement starts to drop when leaving the calibration anchors confidence interval and drops significantly at $\tau_s = 0.8$.

where experts fundamentally disagree on tissue characterization.

Diagnostics. The Localization Sensitivity Analysis (LSA) in Fig. 27 clarifies the nature of this disagreement. The curve exhibits a "Plateau" stability from $\tau_s = 0.2$ to $\tau_s = 0.5$ ($\alpha \approx 0.37$), indicating that radiologists consistently agree on the *anatomical identity* and rough location of nodules. However, the "Cliff" at $\tau_s = 0.6$ (mean $\alpha \to 0.33$) reveals that they disagree on the *morphological extent*. Unlike the sharp vectors of a document layout, biological tissue lacks discrete boundaries, making pixel-perfect segmentation consensus ($\tau_s > 0.6$) achievable only for the most obvious instances.

14.3. Pose Estimation

Finally, we extend KαLOS to the domain of articulated pose estimation using the MARS dataset [38]. This dataset captures the social interactions of mice via a fixed overhead or frontal camera in a standardized cage environment. Unlike bounding boxes or segmentation masks, pose annotations consist of a structured graph of keypoints (*e.g.*, nose, ears,

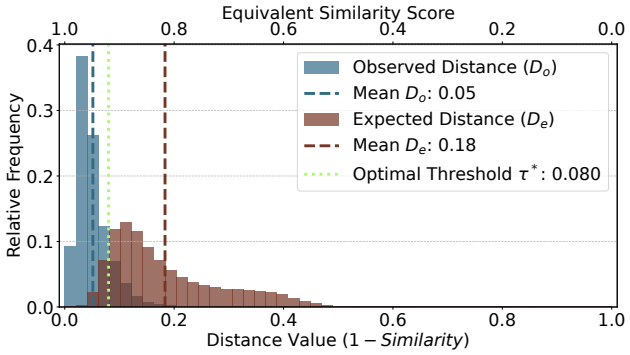

Figure 28. **High Noise Floor.** The distribution of Expected Disagreement (D_e) centers at a remarkably low distance of 0.18. Due to the constrained cage environment and fixed camera, a random mouse pose spatially resembles the target pose, necessitating a strict threshold ($\tau_s^* = 0.92$ or as distance $\tau^* = 0.08$) to statistically isolate the annotator signal (D_o).

hips). This requires a distance metric that captures both the internal structural consistency (pose) and the global placement (location), while being invariant to image resolution.

Distance Function. We use an adapted Image-Normalized Mean Per-Joint Position Error (N-MPJPE). Let $K^A = \{k_1^A, \ldots, k_M^A\}$ be the set of M keypoints provided by rater A, with coordinates normalized to the image dimensions. The distance function is defined as the average error over the union of visible keypoints $J = V_A \cup V_B$:

$$d_{\text{pose}}(K^A, K^B) = 1 - \underbrace{\frac{1}{|J|} \sum_{j \in J} \delta(k_j^A, k_j^B)}_{\text{N-MPJPE}} \tag{30}$$

where the per-joint error δ accounts for both spatial deviation and visibility disagreement:

$$\delta(k_j^A, k_j^B) = \begin{cases} \frac{\|k_j^A - k_j^B\|_2}{\sqrt{2}} & \text{if } j \in V_A \cap V_B \\ 1.0 & \text{if } j \in V_A \oplus V_B \end{cases} \tag{31}$$

For keypoints visible to both raters, we calculate the Euclidean distance normalized by the image diagonal ($\sqrt{2}$ in relative coordinates). For keypoints where visibility is disputed (present in one but not the other), we assign a fixed penalty of 1.0.

Threshold Adaptation and Validation. Validating d_{pose} on MARS reveals a critical insight enabled by our framework (Fig. 28). The distributions of Observed Disagreement (D_o) and Expected Disagreement (D_e) are remarkably close, with means of 0.05 and 0.18 respectively. This proximity reflects the highly constrained nature of the dataset: since the mouse is filmed from a fixed overhead camera in a standard cage [38], a randomly selected mouse (Chance) is spatially likely to be very close to the target mouse. This

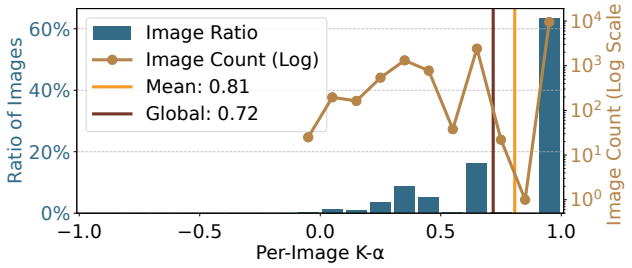

Figure 29. **Per-Image Agreement.** The high median ($\alpha = 1.0$) indicates that for most frames, annotators achieve perfect consensus. However, the distribution tail reveals specific frames where occlusion or motion blur degrades reliability.

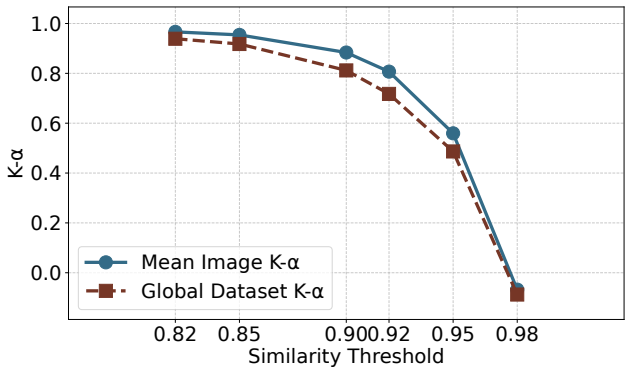

Figure 30. **Localization Sensitivity (LSA).** The metric remains stable across standard thresholds but drops at $\tau_s = 0.95$, revealing the upper bound of inter-annotator precision.

narrow margin necessitates a principled selection of the matching threshold. The Kolmogorov-Smirnov (KS) statistic maximizes at a distance of 0.08 (d_{pose}), corresponding to a similarity threshold of $\tau_s^* = 0.92$ for N-MPJPE (KS = 0.7630). This data-driven calibration confirms that a strict similarity threshold is not merely a preference for precision, but a statistical requirement to separate true matches from environmental coincidence.

Results and Analysis. Using the configuration $\text{K}\alpha\text{LOS}_{\left(d_{\text{pose}},\, \tau=0.08,\, \mathcal{S}=\text{Greedy},\, \psi_{\text{soft}}\right)}$, we obtain a dataset wide mean agreement of $\alpha = 0.8069$. To unpack this result, we examine the distribution of agreement scores across individual images (Fig. 29). The distribution reveals a significant spread with a median of 1.00. The high median indicates that for the majority of frames, annotators achieve near-perfect consensus on pose. However, the tail of the distribution highlights a subset of challenging frames where agreement degrades.

Diagnostics. The Localization Sensitivity analysis (Fig. 30) confirms the logical relationship between strictness and agreement. As the similarity threshold tightens from $\tau_s = 0.82$ to the calibrated anchor $\tau_s^* = 0.92$, α

remains relatively stable ($0.97 \rightarrow 0.81$), indicating robust agreement on the general pose structure. However, at the strict threshold of $\tau_s = 0.98$, agreement drops sharply to -0.07. This inflection point identifies the limit of human precision: while raters agree on the overall pose, sub-pixel precision at the 0.98 similarity level is not achievable even in this constrained domain.