SPOTTING IS NOT INSPECTING: INSPECTBENCH FOR COVERAGE AND RESTRAINT

Anonymous authors
Paper under double-blind review

ABSTRACT

Spotting one defect may flag an image, yet true inspection requires pinpointing them all. We tackle image-only defect inspection with vision–language models (VLMs), challenging a critic model to bound all visible defects or explicitly return an empty set for clean images. We present INSPECTBENCH, a cross-source benchmark designed to evaluate how models navigate the tension between defect *coverage* and false-positive *restraint*. The dataset spans 1,039 cross-reviewed images (321 clean controls) with 1,983 annotated regions across textual and visual defects, evaluated via a deterministic scoring protocol that penalizes duplicate hits and false alarms. Alongside the benchmark, we develop a reference inspection framework that combines content inventory, multi-scale evidence, and candidate verification, together with fine-tuned critic baselines spanning different supervision targets. Evaluating 17 model–procedure configurations across direct prediction and multi-stage inspection reveals a sobering baseline: region recall peaks below 34% at a lenient 0.1 localization threshold, leaving even the top-performing setup unable to touch 46.8% of defect regions. Furthermore, while selective regional review cuts false positives for some models, it backfires on others. These results expose a critical gap between high-level flaw detection and granular visual inspection, proving that evaluating coverage without false-alarm restraint gives a misleading sense of progress. Code, data, and benchmark tools will be released subject to licensing permissions.

1 INTRODUCTION

Localized defect feedback supports image retouching, generator improvement, and visual assessment (Liang et al., 2024; Shen et al., 2026; Wang et al., 2025c; Park et al., 2026). Missed regions can leave defects untreated; false alarms can trigger unnecessary intervention on clean content. Finding one conspicuous defect can flag an image while leaving other defects unexamined.

We study *image-only inspection*. Images received for downstream assessment may arrive without their generation prompts. Judging the image alone also separates visible defects from prompt faithfulness: malformed anatomy or broken glyphs can be assessed without knowing the requested scene or wording. This differs from prompt-conditioned structured defect grounding (Zhang et al., 2026a).

Under one output contract, a critic returns defect regions or an empty set. We measure *coverage* by credited-region recall (R@ t), *restraint* by the fraction of clean images receiving a nonempty prediction (clean FPR), and *localization quality* by geometric region credit. The composite score S combines flawed-image credit and clean-image restraint (Section 4). The lenient 0.1 credit threshold can be met with slight spatial contact; a missed region fails even this permissive check.

INSPECTBENCH contains 1,039 cross-reviewed images, 1,983 annotated regions, and 321 clean controls. Its ownership-based protocol allows one prediction to enclose several nearby defects, but credits each annotated region to at most one prediction. The protocol jointly reports missed regions, weak localization, and clean-image false alarms.

Our baselines span supervision targets and inspection scope. Direct predicts in one call; full review inspects every required inventory unit; Selective uses triage to choose units for local review before candidate verification. We evaluate **11 Direct + 6 multi-stage configurations**; procedure comparisons are made only within a checkpoint. Three findings organize the study:

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

Finding 1: Region recovery remains limited. R@0.1 peaks at 33.8%; the three leading configurations leave roughly 45–50% of regions without overlap (Section 6.3).

Finding 2: Image flagging and recall alone conceal false alarms. A recall-only ranking can prefer a critic that flags roughly four of five clean images over one with much stronger restraint (Section 6.4).

Finding 3: Procedure and supervision coincide with different trade-offs. Selective lowers clean FPR for one of three checkpoints but raises it for two; the Workflow–Decision lineage has lower clean FPR and lower recall (Section 6.5).

Our contributions are:

- **Benchmark:** 1,039 cross-reviewed images with 1,983 annotated regions, 321 clean controls, and multiple regions in 63.4% of flawed images.
- **Region-set protocol:** ownership-based credit supports grouping, penalizes duplicate predictions, and separates no overlap, weak overlap, and credited localization.
- **Baselines and empirical study:** an inspection framework and critics spanning three supervision targets yield 17 configurations, exposing recovery gaps and distinct recall–FPR trade-offs.

2 Related Work

Defect and artifact resources. Cao et al. (2024) organize SynArtifact around an artifact taxonomy. Human-anatomy resources include localized AbHuman (Fang et al., 2024) and HAD (Wang et al., 2024b), and AIGC-HA 1K without boxes (Wang et al., 2025d). ArtiBench (Park et al., 2026) and MagicBench (Wang et al., 2025b) address artifact assessment and fine-grained defect understanding. X-AIGD (Xiao et al., 2026) and SynthScars (Kang et al., 2025) provide categorized pixel-level evidence for synthetic-image detection. These resources differ in annotation scope and purpose; InspectBench applies common image-only annotation and scoring to selected sources.

Structured defects and localized feedback. RichHF-18K (Liang et al., 2024), GenBlemish-27K (Shen et al., 2026), and DiffDoctor (Wang et al., 2025c) connect local defects to feedback or repair; Grounding-IQA (Chen et al., 2026) evaluates grounded quality descriptions and questions. SalArt-VQA separates artifact presence, semantic localization, grounding, and evidence, including abstention on paired references (Sun et al., 2026). SDG (Zhang et al., 2026a) is the closest formulation: it predicts variable-length typed regions from an image and its generation prompt. SDG-Eval includes empty sets and type-wise clean accuracy, alongside class-aware Hungarian box F1 at IoU 0.1 and 0.5. InspectBench instead combines image-only input, ownership-based set credit that permits grouping, and separate reporting of missed regions, weak localization, and clean-image false positives (Table 1).

Table 1: Feature matrix for the closest evaluation protocols. Columns specify input, annotation unit, multi-region scope, clean-control scope, and set credit. Image-only input permits task questions but excludes generation prompts.

Protocol	Image-only	Region-level	Multi-region scope	Clean controls	Set credit
SalArt-VQA	✓	✓	Selected target	Paired reference	Target-wise grounding
SDG-Eval	No	✓	Region set	Type-wise clean	One-to-one matching
InspectBench	✓	✓	Region set	Image-level clean	Group-aware ownership

Complementary evaluation targets. Quality and preference resources assess perceptual judgments or rankings (Li et al., 2024; Wang et al., 2024a; Xu et al., 2023; Kirstain et al., 2023). TIFA, GenEval, T2I-CompBench, and GenAI-Bench evaluate prompt faithfulness and composition (Hu et al., 2023; Ghosh et al., 2023; Huang et al., 2023; Lin et al., 2025). Text and physical-consistency benchmarks add domain-specific references (Zhu et al., 2026; Zhang et al., 2025; Pu et al., 2025). Appendix G groups complementary resources by the information their evaluation requires.

3 THE INSPECTBENCH BENCHMARK

Given only an image, a critic returns a variable-length set of axis-aligned defect predictions, or an empty set. Image-only input limits judgments to visible evidence; region sets represent multiple defects; clean controls measure restraint; and ownership-based credit distinguishes grouping from duplication.

3.1 TASK SCOPE AND SOURCES

The five families are *Rendering*, *Structural*, *Topology*, *Relational-physical*, and *Text integrity*. Text integrity covers textual defects; the other four families cover visual defects. Descriptions record visible evidence. Intended wording, prompt fidelity, aesthetic preference, and image authenticity are outside the task (Appendix A).

Table 2 reports four source groups. GPT-image-2 and SenseNova-U1.5 (Diao et al., 2026) contribute generated images, and natural images provide clean controls. Public-4 pools retained flawed images from ArtiBench, SalArt-VQA, X-AIGD, and TextPecker (Park et al., 2026; Sun et al., 2026; Xiao et al., 2026; Zhu et al., 2026). All selected images undergo the same annotation and cross-review; original source tasks appear in Appendix G. Public-4 contains flawed images only.

Figure 3 (Appendix F) shows five annotated images with 4–7 regions each, mixing text-integrity and visual-defect regions across the examples. In the recruitment poster, small hand and facial regions coexist with a broader band of malformed text and a localized symbol defect. Both small details and larger layout regions must therefore be inspected. The examples include Chinese and Japanese text.

Table 2: Four source groups yield 1,039 images and 1,983 regions. Mean and maximum region counts concern flawed images only. Public-4 comprises ArtiBench, SalArt-VQA, X-AIGD, and TextPecker.

Source group	Images	Flawed	Clean	Regions	Mean	Max.
GPT-image-2	208	97	111	266	2.742	10
Public-4	499	499	0	1,126	2.257	15
SenseNova-U1.5	222	122	100	591	4.844	26
Natural clean	110	0	110	0	Clean-only	
Total	1,039	718	321	1,983	2.762	26

3.2 ANNOTATION AND CROSS-REVIEW

Four annotators perform cross-review, with at least two annotations per candidate image. For two nonempty annotation sets $B^{(a)}$ and $B^{(b)}$, the acceptance criterion requires equal region counts and union-mask IoU above 0.90:

$$|B^{(a)}| = |B^{(b)}|, \qquad \mathrm{IoU}\Big(\bigcup B^{(a)}, \bigcup B^{(b)}\Big) > 0.90. \tag{1}$$

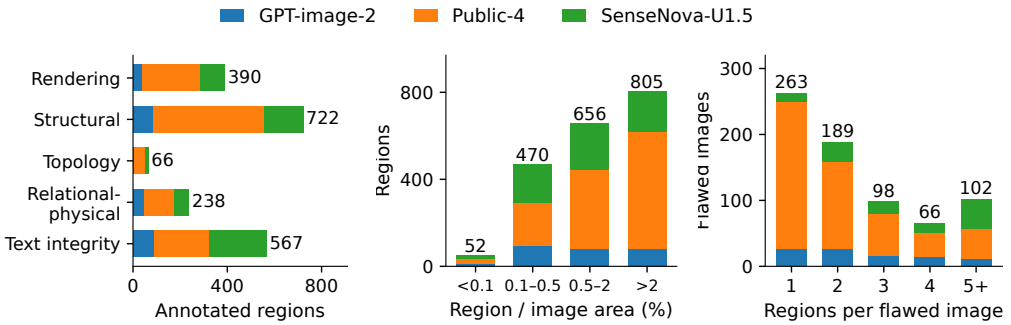
Two empty sets are accepted as a clean label; an empty/nonempty pair enters arbitration. On acceptance, one annotation is retained at random without averaging coordinates. Remaining candidates undergo arbitration and correction; unresolved cases are discarded.

Of 1,429 candidates, 820 pass immediately. The remaining 609 enter arbitration: 219 are corrected and retained, and 390 are discarded, yielding $820 + 219 = 1,039$ retained images. The first-pass acceptance rate is 57.4% and the overall curation yield is 72.7%. Equation (1) is a selection criterion, not an independent agreement statistic on the retained set. Appendix A combines source-specific counts and rates.

3.3 DEFECT MULTIPLICITY AND SPATIAL SCALE

Multiple regions occur in 455 of 718 flawed images. SenseNova-U1.5 contributes 4.844 regions per flawed image on average, versus 2.257 for Public-4, with a maximum of 26. The median region occupies 1.28% of image area, and 59.4% occupy at most 2%. Structural and Text integrity are the

largest families (Figure 1); Text integrity contains 567 regions. These properties characterize the evaluation workload. All results use the pooled image set.



(a) Defect families (b) Region area (c) Defect multiplicity

Figure 1: Multiple small regions create an inspection workload beyond image flagging. (a) Family and (b) area counts each sum to 1,983 regions; (c) multiplicity counts sum to 718 flawed images. Source groups are shared across panels. Figure 3 shows annotated examples.

3.4 Clean Controls and Clean FPR

The 321 controls comprise 111 GPT-image-2, 100 SenseNova-U1.5, and 110 natural images. Clean FPR is the pooled fraction with nonempty predictions on this fixed mixture; within-checkpoint comparisons use identical controls. The controls are not content-matched to flawed images.

4 Region-Set Evaluation Protocol

A prediction may cover several nearby annotated regions, but each region credits only one prediction, so duplicates earn no additional credit. Our deterministic *ownership-based region-set scorer* evaluates original-image pixel coordinates and allows more area slack for small regions.

Geometry and ownership. For prediction P_j and annotated region G_k, containment is $C_{jk} = |P_j \cap G_k|/|G_k|$ and tightness is $\eta_{jk} = |P_j \cap G_k|/|P_j|$. Pair quality raises their Dice overlap *to the power* $\gamma = 0.7$ and applies an area-ratio penalty. Each intersected region chooses the lexicographic maximum of $(C_{jk}, q_{jk}, -j)$: containment, pair quality, then earlier output order. Containment-first ownership gives fully enclosing predictions precedence over partial matches, supporting grouped regions with a deterministic owner. Tightness is recomputed against the exact union of regions owned by a prediction, with grouping tolerance proportional to the square root of their count. Final credit is $s_k \in [0, 1]$; unowned regions receive zero (Appendix B).

The exponent $\gamma < 1$ softens intermediate geometric errors rather than demanding near-exact boxes. For region area fraction f, the tolerance $\tau(f) = 1.25 + 0.35e^{-f/0.01}$ permits up to 1.60-fold area slack for very small regions, approaching 1.25-fold for larger regions; the 0.01 scale makes this allowance decay around one percent of image area. These fixed constants encode a lenient small-region convention.

Image and benchmark scores. For flawed image i, let M_i and N_i count annotated regions and predictions. Let u_{ij} average the credits owned by prediction j, or be zero when it owns none. Define

$$R_i = \frac{\sum_{k=1}^{M_i} s_{ik}}{M_i}, \quad \mathrm{Prec}_i = \frac{\sum_{j=1}^{N_i} u_{ij}}{N_i}, \quad F_i = \frac{2\mathrm{Prec}_i R_i}{\mathrm{Prec}_i + R_i}. \tag{2}$$

Zero denominators give zero. A clean image receives $F_i = e^{-N_i}$, penalizing even the first false prediction while distinguishing repeated alarms. The flawed component Q_+ averages F_i with weights

$\sqrt{M_i}$, giving multi-region images greater influence without letting their weight grow linearly with defect count. The clean component Q_- averages e^{-N_i}. The composite score is

$$S = 100\left(\lambda\sqrt{Q_+} + (1-\lambda)\sqrt{Q_-}\right)^2, \qquad \lambda = 0.7. \tag{3}$$

Here λ prioritizes flawed-image inspection while retaining a clean-image contribution; the square-root mean discourages imbalance between components. Unlike the geometric exponent γ, these choices operate across images. The score satisfies $0 \le S \le 100$, with 100 for perfect predictions and 9 for an always-empty critic. We report S with R@ t, clean FPR, and failure counts. Score differences below the typical 95% interval half-width of approximately 1.8 points are treated as unresolved in our comparisons.

Table 3: Scorer behavior on constructed geometric cases: grouping can receive full credit, while duplicates and clean-image predictions are penalized.

Case	Result
One region, one identical prediction	$\mathrm{Prec}_i = R_i = F_i = 1$
One region, two identical predictions	Credits $(1, 0)$; $F_i = 2/3$
Two adjacent regions exactly tiled by one prediction	$s_1 = s_2 = F_i = 1$
Clean image, one prediction	$F_i = e^{-1} \simeq 0.37$
Always-empty critic	$Q_+ = 0$, $Q_- = 1$, $S = 9$

Thresholds and diagnostics. We report $\mathrm{R@}t = \Pr(s_k \ge t)$ over annotated regions. The threshold applies to *final quality*, not IoU as in SDG-Eval. For one ungrouped region within the area-tolerance band, $s = D^\gamma$ and $\mathrm{IoU} = D/(2-D)$: $s \ge 0.1$ needs only Dice $\simeq 0.037$ (IoU $\simeq 0.019$), whereas $s \ge 0.5$ needs Dice $\simeq 0.371$ (IoU $\simeq 0.228$). These Dice values are lower bounds on the required overlap term, making the low-recall finding conservative about localization precision.

Let Z be the fraction of regions with no intersection with any prediction. The four exhaustive outcomes are

$$\underbrace{Z}_{\text{no overlap}} + \underbrace{1 - Z - \mathrm{R@0.1}}_{\text{overlap, } s<0.1} + \underbrace{\mathrm{R@0.1} - \mathrm{R@0.5}}_{0.1 \le s < 0.5} + \underbrace{\mathrm{R@0.5}}_{s \ge 0.5} = 1. \tag{4}$$

Image sensitivity is $\Pr(N > 0 \mid M > 0)$ and clean FPR is $\Pr(N > 0 \mid M = 0)$. Background false positives on flawed images appear in Appendix E; Section 6.1 specifies failures and intervals.

5 Baselines: Inspection Framework and Fine-Tuned Critics

Supervision and inspection procedure are the two comparison axes, informed by modular artifact analysis and staged retouching (Burgess et al., 2026; Shen et al., 2026). We use *model* for a backbone or endpoint, *checkpoint* for fixed trained weights, *procedure* for an inference policy, and *configuration* for a measured model/checkpoint–procedure pair. A *critic* is the resulting image-to-region-set predictor.

5.1 Supervision and Checkpoint Lineage

Workflow supervision trains planning, tool execution, and visual acceptance; critic-trajectory supervision trains multi-step criticism; decision-and-region supervision trains image decisions, regions, types, and evidence (Table 4). Workflow-SFT records include planning and visual self-checks but no region-level inspection targets, so this checkpoint serves as a procedural initialization. Decision-SFT continues from Workflow-SFT. Appendix C details the Workflow-SFT and Decision-SFT mixtures; Balanced+ and Pool+ are described only by their expanded scale and class balance.

5.2 Inspection Procedures

Figure 2 reads left to right across two rows. *Inventory* and *multi-scale evidence* target missed small regions in multi-defect images; *verification* and *consolidation* target unsupported findings and duplicate predictions. The inventory's Zone, Entity, Text, and Blind labels denote spatial partitions,

Table 4: Six fine-tuned checkpoints span three supervision targets. The Workflow–Decision continuation supports a fixed-procedure lineage comparison.

Checkpoint	Supervision	Training configuration
Workflow-SFT	Workflow trajectories	Qwen3.6-27B; 8,834 originals; 12,312 repeated records
Trajectory-SFT-R1	Critic trajectories	Repeat-1; stratified validation
Trajectory-SFT-R2	Critic trajectories	Repeat-2
Decision-SFT	Decisions + regions	Continues Workflow-SFT; 4,000 train; 800 validation
Decision-SFT-Balanced+	Expanded critic targets	Expanded, class-balanced mixture
Decision-SFT-Pool+	Expanded critic targets	More than 20,000 examples

objects/groups, text regions, and potentially overlooked areas. Each unit is unreviewed, verified clear, suspicious, or uncertain.

The upper row combines global scouts with local multi-scale review (overview, context, detail). The structure scout examines geometry and relations; the rendering scout examines rendering and text. Proposals from these paths form candidate set $\mathcal{V}$. In the lower row, verification checks alternatives such as occlusion, perspective, or style; adjudication accepts only existing candidates ($\mathcal{A} \subseteq \mathcal{V}$); consolidation groups coupled findings while retaining source IDs.

Direct bypasses these modules in one call. *Full-Seq* and *Full-Sched* locally inspect every required inventory unit, with different execution schedules. *Selective* uses triage to select units for local review: high-confidence clear units may bypass it, while suspicious or uncertain units require it. Section 6.5 reports how this changes clean FPR. Scheduling and batching details are given in Table 14. Conservative runs a separate multi-stage implementation with Workflow-SFT under the same output contract (Appendix D.2).

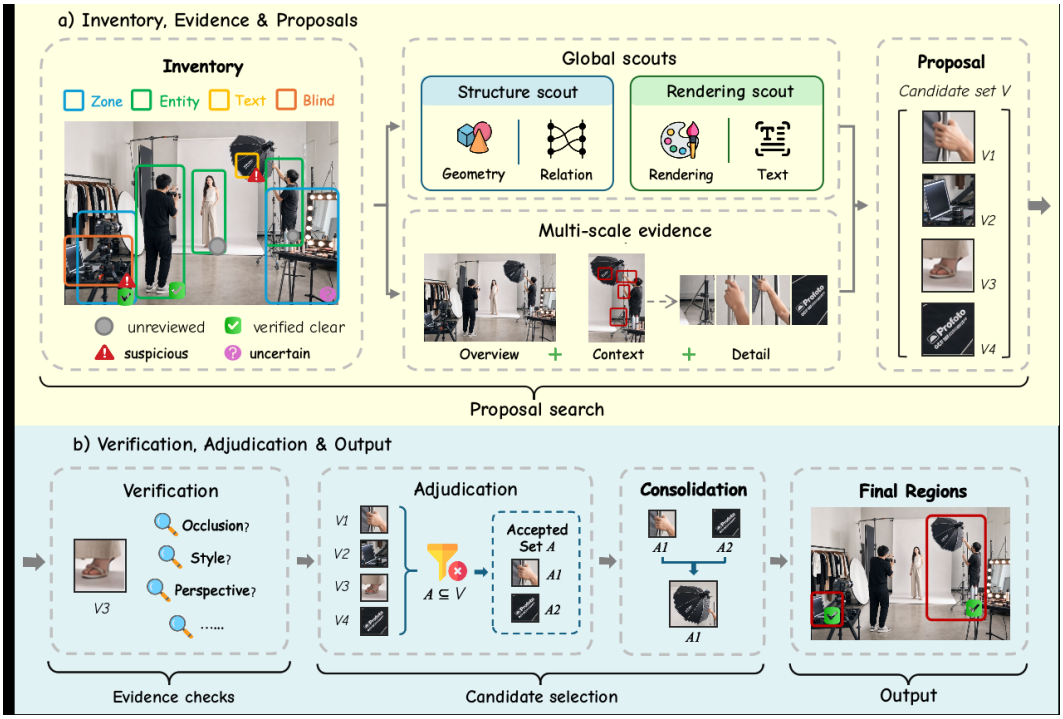


Figure 2: Inspection procedures. Direct maps the image to final regions in one call, bypassing the intermediate stages. Full-Seq and Full-Sched traverse both rows with local review of every required inventory unit. Selective uses triage to restrict local review; all multi-stage routes shown retain candidate verification, adjudication, and consolidation.

Table 5: The 17 evaluated configurations comprise 11 Direct and six multi-stage pairings. Checkmarks denote measured configurations; dashes denote untested combinations. Conservative is evaluated with Workflow-SFT only.

Checkpoint / model	Direct	Full-Seq	Conservative	Full-Sched	Selective
Qwen3.6-27B	✓	–	–	–	–
Qwen3.8-27B	✓	–	–	–	–
Qwen3.7-Plus	✓	–	–	–	–
GPT-5.6-sol	✓	–	–	–	–
Gemini-3.5-Flash	✓	–	–	–	–
Workflow-SFT	✓	✓	✓	✓	✓
Trajectory-SFT-R1	✓	–	–	–	–
Trajectory-SFT-R2	✓	–	–	–	✓
Decision-SFT	✓	–	–	–	✓
Decision-SFT-Balanced+	✓	–	–	–	–
Decision-SFT-Pool+	✓	–	–	–	–

Table 5 supports within-checkpoint procedure comparisons and fixed-procedure checkpoint comparisons. Direct versus Selective is available for three checkpoints; full-review comparisons use Workflow-SFT.

6 Experiments

6.1 Setup

All 17 configurations use the same 1,039 images, output contract, and scorer. The base/commercial identifiers are Qwen3.6-27B, Qwen3.8-27B, Qwen3.7-Plus, GPT-5.6-sol, and Gemini-3.5-Flash; checkpoint lineage appears in Table 4.

Predictions use normalized $[x, y, w, h]$ coordinates on a 0–1,000 scale, converted to original-image pixels; corner-form and crop-relative outputs are transformed first (Appendix B). Error, missing, and absent records are scored as empty without dropping images: failed flawed images receive $F_i = 0$, failed clean images $F_i = 1$. Intervals use 2,000 stratified image-bootstrap replicates, resampling flawed and clean images separately and taking the 2.5th/97.5th percentiles. Diagnostics concern final outputs.

6.2 Main Results

Composite scores summarize different recall–FPR operating points. Decision-SFT/Direct uses one call and has substantially lower clean FPR than the full-review configurations, which recover more regions. Its S interval overlaps Full-Sched's (Table 6). The three highest-S point estimates have overlapping intervals and differ by no more than 0.5 points.

Six configurations are nondominated in clean FPR and R@0.1. In increasing clean-FPR order, they are Pool+, Balanced+, Decision-SFT/Direct, Decision-SFT/Selective, Workflow-SFT/Full-Sched, and Workflow-SFT/Full-Seq. Each remaining configuration has another with no higher clean FPR and no lower recall, and at least one strict improvement. The always-empty critic anchors $(0, 0)$. Consistent with Finding 1, even the lenient recall threshold remains difficult, and the greatest spatial reach covers only 56.0% of annotated regions (Appendix E).

6.3 Where Region Credit Is Lost

Overlap is not credit under this protocol. For Decision-SFT, Selective reduces the no-overlap fraction by 12.4 points but increases R@0.1 by only 5.8 points (Table 7). The difference accumulates as weak overlap: more regions are touched without reaching the final-quality threshold.

The three leading configurations—Full-Seq, Full-Sched, and Decision-SFT/Selective—leave 45.4–50.4% of regions without any overlapping prediction (Tables 7 and 16). Missing overlap and weak localization identify distinct failures.

Table 6: Recall–FPR operating points and composite score S [95% CI]. Recall, sensitivity (Sens.), and clean FPR are percentages; bold marks column-best point estimates. #fail counts failed/missing image outputs, scored as empty. Maximum S credit from failed clean outputs is 1.448 points for Full-Seq, 0.390 for Conservative, 0.129 for Trajectory-SFT-R2/Selective, 0.064 for Full-Sched, and 0.059 for Decision-SFT/Selective (Appendix B.5).

Model / procedure	$S \uparrow$ [95% CI]	R@0.1↑	R@0.5↑	Sens.↑	Clean FPR↓	#fail
Direct: base / commercial						
Qwen3.6-27B	23.4 [21.8, 24.9]	30.7	23.1	**99.6**	95.6	0
Qwen3.8-27B	23.9 [22.4, 25.5]	24.0	17.6	98.1	89.4	0
Qwen3.7-Plus	19.7 [18.1, 21.2]	26.5	18.3	98.1	80.1	0
GPT-5.6-sol	20.2 [18.6, 21.7]	8.0	6.0	50.6	16.2	0
Gemini-3.5-Flash	16.3 [15.0, 17.6]	11.3	3.6	93.7	48.6	0
Direct: fine-tuned						
Workflow-SFT	28.5 [26.7, 30.3]	24.4	18.2	71.7	35.5	0
Trajectory-SFT-R1	24.5 [22.8, 26.2]	33.3	23.9	97.2	81.9	0
Trajectory-SFT-R2	25.4 [23.7, 27.1]	33.2	23.9	97.4	81.6	0
Decision-SFT	30.4 [28.6, 32.2]	23.3	16.5	61.0	10.9	0
Decision-SFT-Balanced+	22.4 [20.8, 24.1]	8.6	6.3	30.8	4.4	0
Decision-SFT-Pool+	19.2 [17.7, 20.6]	5.1	3.6	19.4	**2.5**	0
Multi-stage inspection						
Workflow-SFT / Full-Seq	31.5 [29.8, 33.3]	**33.8**	23.5	83.7	44.2	22
Workflow-SFT / Conservative	28.7 [27.0, 30.5]	23.0	15.7	56.3	25.2	7
Workflow-SFT / Full-Sched	**31.9** [30.1, 33.7]	33.4	**24.0**	83.3	38.6	1
Workflow-SFT / Selective	29.4 [27.7, 31.2]	31.0	21.1	87.3	46.7	0
Trajectory-SFT-R2 / Selective	29.1 [27.3, 30.8]	32.4	21.4	84.1	48.3	2
Decision-SFT / Selective	31.4 [29.7, 33.2]	29.1	19.3	77.7	26.2	1

Table 7: Additional spatial reach need not earn region credit. The four outcome fractions (%) in Eq. (4) sum to 100 up to rounding; weak overlap means an intersection with $s < 0.1$.

Model / procedure	No overlap	Weak overlap	$0.1 \leq s < 0.5$	$s \geq 0.5$
Workflow-SFT / Full-Seq	45.4	20.8	10.3	23.5
Workflow-SFT / Full-Sched	46.8	19.9	9.4	24.0
Decision-SFT / Direct	62.8	14.0	6.7	16.5
Decision-SFT / Selective	50.4	20.6	9.8	19.3

6.4 Image-Level Decisions versus Region Inspection

High-level detection is an image-level flag, $N > 0$**.** Five configurations have sensitivity at least 97%, yet their R@0.1 is at most 33.3%. These flags are nearly indiscriminate (clean FPR 80–96%); the gap is between flagging an image and localizing its defects.

A recall-only ranking can prefer frequent false alarms. Ordering the tabulated point estimates by R@0.1 places Trajectory-SFT-R1 third and Decision-SFT/Direct twelfth; by S they reverse to tenth and fourth. The former flags 97.2% of flawed and 81.9% of clean images; the latter flags 61.0% and 10.9%, respectively (Table 6). Recall-only reporting would thus prefer the critic that raises a false alarm on roughly four of five clean images. This reversal illustrates Finding 2: S accounts for restraint, and the recall–FPR pair makes the underlying trade-off explicit.

Restraint alone is also insufficient. Balanced+ and Pool+ flag only 30.8% and 19.4% of flawed images, with R@0.1 of 8.6% and 5.1%, respectively. They lie toward the always-empty end of the trade-off.

6.5 Procedure and Supervision Trade-offs

Selective lowers clean FPR for one of three checkpoints (Trajectory-SFT-R2, -33.3 points) **and raises it for the other two** (Workflow-SFT, $+11.2$; Decision-SFT, $+15.3$). The corresponding S point changes are $+3.7$, $+0.9$, and $+1.0$; only the first exceeds the reporting margin in Section 4.

The clean-FPR changes vary with the starting point in this comparison ($n = 3$). Selective's clean FPRs are 48.3/46.7/26.2%, more concentrated than Direct's 81.6/35.5/10.9%. The only reduction occurs at the highest starting FPR.

At fixed Workflow-SFT, Full-Sched has higher recall and modestly higher clean FPR than Direct, while Full-Seq reaches the highest recall with 22 failed image outputs (Table 6). Under Direct, the Qwen3.6-27B→Workflow-SFT→Decision-SFT lineage instead loses recall monotonically, 30.7→24.4→23.3%. Its flawed-image score component also falls, 21.15→16.16→14.21; the rise in S is driven by the clean component, 28.99→70.87→91.92 (Table 15). At fixed Selective, Decision-SFT similarly has lower recall and lower FPR than its Workflow-SFT parent. Finding 3 is that multi-stage review coincides with higher recall (Full-Sched versus Direct at Workflow-SFT), while the Workflow–Decision lineage coincides with lower clean FPR and lower recall; neither comparison improves both axes.

7 LIMITATIONS

Data. Acceptance under Eq. (1) compares region counts and union masks, not per-region or per-family agreement; 42.6% of candidates required arbitration. A clean label means annotators found no defect, so some model false positives may be genuine defects. Clean images come from three sources while Public-4 contributes flawed images only; clean FPR uses this fixed mixture without per-source rates. Results concern the retained source and script mixture, with no language-balanced or out-of-distribution evaluation.

Training and procedures. Decision-SFT's Wikimedia/Pinterest negatives are construction-assigned, and natural images are 34.3% of the controls, so domain alignment may contribute to its low FPR. Procedures differ in inference budget, the pairing matrix is incomplete, and training ingredients vary jointly, so comparisons are descriptive.

Protocol. Containment-first ownership can favor a broad box over a tighter partial one; tolerance constants and aggregation weights are fixed conventions.

8 CONCLUSION

INSPECTBENCH combines flawed-image credit and clean-image restraint in the composite score S, with diagnostics that separate touching a defect, earning region credit, and avoiding false alarms. At fixed Workflow-SFT, Full-Sched has higher recall and higher clean FPR than Direct; the Workflow–Decision lineage coincides with lower clean FPR and lower recall. Neither comparison improves both axes. We suggest reporting S together with recall–FPR pairs on clean-inclusive benchmarks, failure counts, and localization diagnostics: image-level flagging and recall alone can favor critics that frequently raise false alarms on clean content.

Ethics Statement

Four annotators performed annotation and cross-review, with at least two annotations per candidate. The common annotation guide defines the five defect families, requires visible and independently localizable evidence, and distinguishes coupled from separate symptoms (Appendix A). Disagreements enter arbitration and unresolved candidates are discarded. We report acceptance and curation outcomes without identifying annotators.

Image redistribution follows source-specific permissions. Released assets retain source attribution; rights-restricted images are represented by permissible annotations, identifiers, or retrieval instructions.

Reproducibility Statement

The supplementary material contains the deterministic scorer, 24 synthetic conformance tests, annotation and review statistics, and all 17 configuration-level result records. Appendix B specifies geometry, ownership, grouping, failure handling, aggregation, and bootstrap computation; Appendix D gives the multi-stage procedure contracts. Per-image prediction JSONL is the scorer's input format, with synthetic examples supplied. Public data and tool releases follow the licensing policy above.

Anonymous benchmark resources and related code:
https://anonymous.4open.science/r/InspectBench-D407.

AI Use Statement

We used generative AI tools for writing polishing, manuscript organization, retrieving related work, language editing, citation checking, and writing code for conformance tests and numerical checks. The study also utilizes model-generated content, including synthetic datasets, generated images, model-produced workflow trajectories and critiques, and model predictions, as detailed in Sections 3 and 5 and Appendix C. Generative AI was not used for designing the core methodology or experiments, formulating research claims, or producing primary model-evaluation results; translation and dataset cleaning were not applicable. Benchmark annotation and cross-review were conducted entirely by human annotators (Appendix A). All AI-assisted text, claims, code, and artifacts were thoroughly verified by the authors, who take full responsibility for the final content of this work.

References

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD—a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. doi: 10.1109/CVPR.2019.00982. URL https://www.mvtec.com/research-teaching/datasets/mvtec-ad.

Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *International Journal of Computer Vision*, 2022. doi: 10.1007/s11263-022-01578-9. URL https://www.mvtec.com/research-teaching/datasets/mvtec-loco-ad.

Daniel Bolya, Sean Foley, James Hays, and Judy Hoffman. TIDE: A general toolbox for identifying object detection errors. In *European Conference on Computer Vision*, 2020. URL https://arxiv.org/abs/2008.08115.

James Burgess, Rameen Abdal, Dan Stoddart, Sergey Tulyakov, Serena Yeung-Levy, and Kuan-Chieh Jackson Wang. ArtifactLens: Hundreds of labels are enough for artifact detection with VLMs. *arXiv preprint arXiv:2602.09475*, 2026. doi: 10.48550/arXiv.2602.09475. URL https://arxiv.org/abs/2602.09475.

Bin Cao, Jianhao Yuan, Yexin Liu, Jian Li, Shuyang Sun, Jing Liu, and Bo Zhao. SynArtifact: Classifying and alleviating artifacts in synthetic images via vision-language model. *arXiv preprint arXiv:2402.18068*, 2024. doi: 10.48550/arXiv.2402.18068. URL https://arxiv.org/abs/2402.18068.

Zheng Chen, Xun Zhang, Wenbo Li, Renjing Pei, Fenglong Song, Xiongkuo Min, Xiaohong Liu, Xin Yuan, Yong Guo, and Yulun Zhang. Grounding-IQA: Grounding multimodal language model for image quality assessment. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2411.17237.

Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.18235.

Haiwen Diao, Jiahao Wang, Chenjing Ding, Hanming Deng, Jiangnan Chen, Ruixi Zhang, Ruohui Wang, Wenwen Tong, Xiangyu Fan, Yubo Wang, Yue Zhu, Yuwei Niu, Zhengqi Bai, Zhiqian Lin, Zhitao Yang, Zhongang Cai, Bo Yang, Chen Feng, Chengguang Lv, Guangjia Liu, Guanlin Wang, Hanyu Zhang, Haojia Yu, Hongcan Xiao, Hongli Wang, Huan Wu, Huaping Zhong, Jian Fang, Jianan Fan, Jiaqi Li, Jiefan Lu, Jing Zuo, Jingcheng Ni, Junxiang Xu, Linjun Dai, Mutian Xu, Peishen Yan, Penghao Wu, Ruijie Mao, Ruisi Wang, Shihao Bai, Shuang Yang, Shuya Yang, Shuyan Zheng, Silei Wu, Siying Li, Tao Chu, Tianbo Zhong, Tongxi Zhou, Weichao Luo, Weichen Fan, Wenhao Jia, Wenjie Gao, Xiangli Kong, Yan Li, Yang Yong, Zimo Wen, Zixuan Qian, Wenxiu Sun, Ruihao Gong, Quan Wang, Lewei Lu, Lei Yang, Ziwei Liu, and Dahua Lin. SenseNova-U1.5: Towards native unified visual intelligence. *arXiv preprint arXiv:2609.11929*, 2026. doi: 10.48550/arXiv.2609.11929. URL https://arxiv.org/abs/2609.11929.

Guian Fang, Wenbiao Yan, Yuanfan Guo, Jianhua Han, Zutao Jiang, Hang Xu, Shengcai Liao, and Xiaodan Liang. HumanRefiner: Benchmarking abnormal human generation and refining with coarse-to-fine pose-reversible guidance. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2407.06937.

Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2310.11513.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2310.14566.

Shengbin Guo, Shaokang He, Chaoyue Meng, Shengpeng Xiao, Xunzhi Xiang, Shaofeng Zhang, and Qi Fan. PhyEditBench: A real-world multi-stage benchmark for physics-aware image editing. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2606.26551.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2021. URL https://arxiv.org/abs/2104.08718.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. ELLA: Equip diffusion models with LLM for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. doi: 10.48550/arXiv.2403.05135. URL https://arxiv.org/abs/2403.05135.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2303.11897.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2307.06350v2.

Hengrui Kang, Siwei Wen, Zichen Wen, Junyan Ye, Weijia Li, Peilin Feng, Baichuan Zhou, Bin Wang, Dahua Lin, Linfeng Zhang, and Conghui He. LEGION: Learning to ground and explain for

synthetic image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Kang_LEGION_Learning_to_Ground_and_Explain_for_Synthetic_Image_Detection_ICCV_2025_paper.html.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-Pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.01569.

Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. AGIQA-3K: An open database for AI-generated image quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8):6833–6846, 2024. doi: 10.1109/TCSVT.2023.3319020. URL https://ieeexplore.ieee.org/document/10262331/.

Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, Junjie Ke, Krishnamurthy Dj Dvijotham, Katie Collins, Yiwen Luo, Yang Li, Kai J Kohlhoff, Deepak Ramachandran, and Vidhya Navalpakkam. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2312.10240.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *Computer Vision – ECCV 2024*, volume 15067 of *Lecture Notes in Computer Science*, pp. 366–384. Springer, 2025. doi: 10.1007/978-3-031-72673-6_20. URL https://link.springer.com/chapter/10.1007/978-3-031-72673-6_20.

Samin Mahdizadeh Sani, Max Ku, Nima Jamali, Matina Sani, Paria Khoshtab, Wei-Chieh Sun, Parnian Fazel, Zhi Rui Tam, Thomas Chong, Edisy Kin Wai Chan, Donald Tsang, Chiao-Wei Hsu, Ting Lam, Ho Ng, Chiafeng Chu, Chak-Wing Mak, Keming Wu, Hiu-Tung Wong, Yik Ho, Chi Ruan, Zhuofeng Li, I-Sheng Fang, Shih-Ying Yeh, Ho Kei Cheng, Ping Nie, and Wenhu Chen. ImagenWorld: Stress-testing image generation models with explainable human evaluation on open-ended real-world tasks. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/458fa8ee331566383d8e74bdb647f829-Abstract-Conference.html.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019. doi: 10.1145/3287560.3287596. URL https://arxiv.org/abs/1810.03993.

Jaehyun Park, Minyoung Ahn, Minkyu Kim, Jonghyun Lee, Jae-Gil Lee, and Dongmin Park. See and fix the flaws: Enabling VLMs and diffusion models to comprehend visual artifacts via agentic data synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2602.20951.

Yuandong Pu, Le Zhuo, Songhao Han, Jinbo Xing, Kaiwen Zhu, Shuo Cao, Bin Fu, Si Liu, Hongsheng Li, Yu Qiao, Wenlong Zhang, Xi Chen, and Yihao Liu. PICABench: How far are we from physically realistic image editing? *arXiv preprint arXiv:2510.17681*, 2025. doi: 10.48550/arXiv.2510.17681. URL https://arxiv.org/abs/2510.17681.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. URL https://arxiv.org/abs/1902.09630.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual CoT: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2403.16999.

Shaocheng Shen, Jianfeng Liang, Chunlei Cai, Cong Geng, Huiyu Duan, Xiaoyun Zhang, Qiang Hu, and Guangtao Zhai. Agentic retoucher for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2601.02046.

Xiaoxiao Sun, Ruotian Zhang, Junzhe Huang, James Burgess, and Serena Yeung-Levy. SalArt-VQA: Diagnosing whether VLMs understand salient artifacts in generated images. *arXiv preprint arXiv:2606.12671*, 2026. doi: 10.48550/arXiv.2606.12671. URL https://arxiv.org/abs/2606.12671.

Alex Jinpeng Wang, Dongxing Mao, Jiawei Zhang, Weiming Han, Zhuobai Dong, Linjie Li, Yiqi Lin, Zhengyuan Yang, Libo Qin, Fuwei Zhang, Lijuan Wang, and Min Li. TextAtlas5M: A large-scale dataset for long and structured text image generation. *arXiv preprint arXiv:2502.07870*, 2025a. doi: 10.48550/arXiv.2502.07870. URL https://arxiv.org/abs/2502.07870v2. Accepted to ICML 2026.

Jia Wang, Jie Hu, Xiaoqi Ma, Hanghang Ma, Yanbing Zeng, and Xiaoming Wei. MagicMirror: A large-scale dataset and benchmark for fine-grained artifacts assessment in text-to-image generation. *arXiv preprint arXiv:2509.10260*, 2025b. doi: 10.48550/arXiv.2509.10260. URL https://arxiv.org/abs/2509.10260.

Jiarui Wang, Huiyu Duan, Jing Liu, Shi Chen, Xiongkuo Min, and Guangtao Zhai. AIGCIQA2023: A large-scale image quality assessment database for AI generated images: from the perspectives of quality, authenticity and correspondence. In *Artificial Intelligence – CICAI 2023*, volume 14474 of *Lecture Notes in Computer Science*, pp. 46–57. Springer, 2024a. doi: 10.1007/978-981-99-9119-8_5. URL https://link.springer.com/chapter/10.1007/978-981-99-9119-8_5.

Kaihong Wang, Lingzhi Zhang, and Jianming Zhang. Detecting human artifacts from text-to-image models. *arXiv preprint arXiv:2411.13842*, 2024b. doi: 10.48550/arXiv.2411.13842. URL https://arxiv.org/abs/2411.13842.

Yiyang Wang, Xi Chen, Xiaogang Xu, Sihui Ji, Yu Liu, Yujun Shen, and Hengshuang Zhao. DiffDoctor: Diagnosing image diffusion models before treating. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025c. URL https://arxiv.org/abs/2501.12382.

Zeqing Wang, Qingyang Ma, Wentao Wan, Haojie Li, Keze Wang, and Yonghong Tian. Is this generated person existed in real-world? fine-grained detecting and calibrating abnormal human-body. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21226–21237, 2025d. URL https://arxiv.org/abs/2411.14205.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. Q-Bench: A benchmark for general-purpose foundation models on low-level vision. In *International Conference on Learning Representations*, 2024a. URL https://arxiv.org/abs/2309.14181.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, Geng Xue, Wenxiu Sun, Qiong Yan, and Weisi Lin. Q-Instruct: Improving low-level visual abilities for multi-modality foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b. URL https://arxiv.org/abs/2311.06783.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. In *Proceedings of the International Conference on Machine Learning*, 2024c. URL https://arxiv.org/abs/2312.17090.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2312.14135.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. doi: 10.48550/arXiv.2306.09341. URL https://arxiv.org/abs/2306.09341.

Yao Xiao, Weiyan Chen, Jiahao Chen, Zijie Cao, Weijian Deng, Binbin Yang, ZiYi Dong, Xiangyang Ji, Wei Ke, Pengxu Wei, and Liang Lin. Unveiling perceptual artifacts: A fine-grained benchmark for interpretable AI-generated image detection. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/65b4e84b66ed049f8066919a803e3942-Abstract-Conference.html.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2304.05977.

Zhiyuan You, Zheyuan Li, Jinjin Gu, Zhenfei Yin, Tianfan Xue, and Chao Dong. Depicting beyond scores: Advancing image quality assessment through multi-modal language models. In *European Conference on Computer Vision*, 2024. doi: 10.1007/978-3-031-72970-6_15. URL https://arxiv.org/abs/2312.08962.

Huaisong Zhang, Hao Yu, Yuxuan Zhang, Jiahe Wang, Xinrui Chen, Haoxiang Cao, Feng Lu, Wendong Zhang, Changqian Yu, and Chun Yuan. Where, what, why, and importance: Structured defect grounding for text-to-image feedback. *arXiv preprint arXiv:2606.06113*, 2026a. doi: 10.48550/arXiv.2606.06113. URL https://arxiv.org/abs/2606.06113.

Peirong Zhang, Haowei Xu, Jiaxin Zhang, Xuhan Zheng, Guitao Xu, Yuyi Zhang, Junle Liu, Zhenhua Yang, Wei Zhou, and Lianwen Jin. OCRGenBench: A comprehensive benchmark for evaluating OCR generative capabilities. *arXiv preprint arXiv:2507.15085*, 2025. doi: 10.48550/arXiv.2507.15085. URL https://arxiv.org/abs/2507.15085.

Tong Zhang, Honglin Lin, Zhou Liu, Chong Chen, and Wentao Zhang. SciFlow-Bench: Evaluating structure-aware scientific diagram generation via inverse parsing. *arXiv preprint arXiv:2602.09809*, 2026b. doi: 10.48550/arXiv.2602.09809. URL https://arxiv.org/abs/2602.09809.

Qinye Zhou, Jun Zheng, Yongchao Du, Yuan Wang, Zhengrui Chen, Zuan Gao, Taihang Hu, Chao Lin, Yefeng Shen, Xingjian Wang, Zhao Wang, Zhengtao Wu, Xiaoli Xu, Zhengze Xu, Hao Yan, Denghui Yang, Yuhang Yu, Huayu Zhang, Mingzhou Zhang, and Mengting Chen. CPI-Bench: A comprehensive, practical and intelligent benchmark for real-world image editing. *arXiv preprint arXiv:2608.14546*, 2026. doi: 10.48550/arXiv.2608.14546. URL https://arxiv.org/abs/2608.14546.

Hanshen Zhu, Yuliang Liu, Xuecheng Wu, An-Lan Wang, Hao Feng, Dingkang Yang, Chao Feng, Can Huang, Jingqun Tang, and Xiang Bai. TextPecker: Rewarding structural anomaly quantification for enhancing visual text rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2602.20903.

A Annotation Protocol and Additional Dataset Statistics

A.1 Scope and Acceptance

Annotators mark a region when a defect is visually supported and independently localizable. Coupled symptoms can share one region; separate descriptions of the same visible problem do not automatically require separate regions. Routine defocus, compression, or artistic style is marked only with direct evidence of an unintended visible defect. Text integrity concerns damaged glyph structure.

Table 8: Five families organize the annotated evidence. Predicted family labels do not enter the composite score.

Family	Visible evidence
Rendering	Local texture, boundary, color, or rendering discontinuity
Structural	Malformed entity shape, anatomy, or component structure
Topology	Broken, fused, or inconsistent connectivity between parts
Relational-physical	Inconsistent interaction, support, geometry, or physical relation
Text integrity	Broken, fused, incomplete, or malformed glyph structure

Equation (1) requires equal region counts and union-mask IoU above 0.90. Retained coordinates come from one accepted annotation, chosen at random after acceptance, or from an arbitrated correction.

A.2 Acceptance and Arbitration by Source

Table 9 combines source-level review counts and rates. Across all 1,429 candidates, 609 require arbitration (42.6%); 219 of those are corrected and retained (36.0%), and 390 are discarded (64.0%).

Table 9: Cross-review counts and rates. Parentheses give percentages of each source's candidate count. Yield is the retained fraction.

Source group	Candidates	First-pass	Corrected	Discarded	Retained	Yield
GPT-image-2	242	182 (75.2%)	26 (10.7%)	34 (14.0%)	208	86.0%
Public-4	712	384 (53.9%)	115 (16.2%)	213 (29.9%)	499	70.1%
SenseNova-U1.5	350	152 (43.4%)	70 (20.0%)	128 (36.6%)	222	63.4%
Natural clean	125	102 (81.6%)	8 (6.4%)	15 (12.0%)	110	88.0%
Total	1,429	820 (57.4%)	219 (15.3%)	390 (27.3%)	1,039	72.7%

Within Public-4, first-pass acceptance is 303/504 (60.1%) in the lower-density workload group and 81/208 (38.9%) in the dense workload group.

A.3 Fine-Grained Multiplicity and Spatial Scale

Table 10 expands Figure 1(c) into exact counts; omitted counts have zero images. Family and area-bin counts are included in the aggregate data.

The mean/median region area fractions are 2.11%/0.88% for GPT-image-2, 4.95%/1.85% for Public-4, and 3.09%/0.86% for SenseNova-U1.5. The pooled mean/median is 4.01%/1.28%. These are per-region statistics, so images with more annotated regions contribute more observations.

Table 10: Exact region-count distribution among flawed images. Source columns sum to 97, 499, and 122 images; the total is 718.

Regions/image	GPT-image-2	Public-4	SenseNova-U1.5	Total
1	27	223	13	263
2	27	132	30	189
3	16	63	19	98
4	15	36	15	66
5	3	14	9	26
6	5	15	5	25
7	2	4	8	14
8	1	3	7	11
9	0	4	2	6
10	1	3	2	6
11	0	0	3	3
12	0	0	1	1
13	0	1	3	4
14	0	0	1	1
15	0	1	0	1
16	0	0	1	1
17	0	0	1	1
19	0	0	1	1
26	0	0	1	1

B Complete Scorer Specification

B.1 Inputs and Coordinate Conventions

Annotation JSONL uses unique `sample_id`, original image dimensions, and an `annotation` object containing `no_error` and `boxes`. Each box is `bbox_xywh` in absolute pixels. A clean image must have `no_error=true` and an empty box list; a flawed image must have a nonempty list. Annotated regions must be in bounds and have positive width and height.

Prediction JSONL uses the same image ID, a `status`, and a list of `bbox_1000_xywh` boxes. Convert normalized $[x, y, w, h]$ to pixels by multiplying horizontal coordinates by image width/1,000 and vertical coordinates by image height/1,000. When an inference procedure emits $[x_1, y_1, x_2, y_2]$, its adapter first converts to $[x_1, y_1, x_2 - x_1, y_2 - y_1]$. Crop-relative coordinates must be mapped to the original image before this conversion. Geometric areas are continuous rectangle areas; there is no inclusive-pixel "$+1$" term.

The scorer recognizes four artifact families and two text families. InspectBench uses the four artifact families and text integrity; it contains no requested-text-fidelity targets. Unknown predicted families do not remove their boxes from geometric scoring. Descriptions are retained for inspection but do not enter a score.

B.2 Pair Quality and Deterministic Ownership

For prediction P_j, annotated region G_k, and image area A_I, compute

$$I_{jk} = |P_j \cap G_k|, \qquad C_{jk} = I_{jk}/|G_k|, \quad \eta_{jk} = I_{jk}/|P_j|, \tag{5}$$

$$D_{jk} = \begin{cases} 2C_{jk}\eta_{jk}/(C_{jk} + \eta_{jk}), & C_{jk} + \eta_{jk} > 0, \\ 0, & \text{otherwise}, \end{cases} \tag{6}$$

then

$$f_k = |G_k|/A_I, \qquad r_{jk} = |P_j|/|G_k|, \tag{7}$$

$$\tau(f) = 1.25 + 0.35\exp(-f/0.01), \tag{8}$$

$$e_{jk} = \max\{0, |\log r_{jk}| - \log \tau(f_k)\}, \tag{9}$$

$$\psi_{jk} = \exp(-4e_{jk}^2), \qquad q_{jk} = D_{jk}^{\gamma}\psi_{jk}. \tag{10}$$

The symmetric tolerance band is $[1/\tau, \tau]$. Its width decreases smoothly from 1.60 for very small area fractions toward 1.25 for large regions; at $f = 0.01$, $\tau \simeq 1.379$. The exponent $\gamma = 0.7$ gives intermediate Dice values more credit than a linear Dice term, while the area penalty suppresses substantial over- or under-enclosure outside the tolerance band.

Define $J_k = \{j : I_{jk} > 0\}$. If J_k is nonempty, the owner is

$$a(k) = \operatorname*{lex\,argmax}_{j \in J_k}(C_{jk}, q_{jk}, -j). \tag{11}$$

The comparison is lexicographic: the second key is considered only when containment is equal, and the third only when both earlier keys are equal. The first prediction wins an otherwise exact tie. An unoverlapped annotated region has no owner. Ownership is determined from one-to-one pair statistics before merged scoring and is not recomputed afterward.

This ordering prioritizes containment, giving fully enclosing predictions precedence over partial matches. Pair quality and output order resolve ties, and the owned region set determines the subsequent union calculation.

B.3 Merged-Region Recalculation

For each prediction, let $O_j = \{k : a(k) = j\}$, $K_j = |O_j|$, and

$$U_j = \bigcup_{k \in O_j} G_k. \tag{12}$$

For $K_j > 0$, compute

$$\eta_j^{\cup} = \frac{|P_j \cap U_j|}{|P_j|}, \qquad r_j^{\cup} = \frac{|P_j|}{|U_j|}, \qquad \tau_j^{\cup} = \tau\left(\frac{|U_j|}{A_I}\right)\sqrt{K_j}, \tag{13}$$

$$e_j^{\cup} = \max\{0, |\log r_j^{\cup}| - \log \tau_j^{\cup}\}, \quad \psi_j^{\cup} = \exp[-4(e_j^{\cup})^2], \tag{14}$$

$$D_k^{\cup} = \frac{2C_{jk}\eta_j^{\cup}}{C_{jk} + \eta_j^{\cup}}, \qquad s_k = (D_k^{\cup})^{\gamma}\psi_j^{\cup} \quad (k \in O_j). \tag{15}$$

Unowned regions have $s_k = 0$. The $K_j = 1$ case reduces to pair scoring. A group uses the union of its rectangles, not their enclosing rectangle or the sum of their areas. An exact sweep over x-boundaries, with merged y-intervals in each strip, computes union area; intersecting each owned rectangle with P_j gives the corresponding intersection-union area. Overlapping annotated regions are not double-counted in these areas.

The square-root factor is a grouping-tolerance heuristic. It accommodates some enclosure slack when several nearby defects share a prediction, while the union tightness term penalizes uncovered gaps between them.

B.4 Image Scores and Aggregation

Prediction credit is the mean of the qualities it owns:

$$u_j = \begin{cases} K_j^{-1} \sum_{k \in O_j} s_k, & K_j > 0, \\ 0, & K_j = 0. \end{cases} \tag{16}$$

For a flawed image, use Eq. (2); for a clean image use $F_i = \exp(-N_i)$. Then

$$Q_+ = \frac{\sum_{i:M_i>0} \sqrt{M_i} F_i}{\sum_{i:M_i>0} \sqrt{M_i}}, \qquad Q_- = \frac{1}{n^-} \sum_{i:M_i=0} \exp(-N_i), \tag{17}$$

and combine them with Eq. (3). The component weight $\lambda = 0.7$, component power 0.5, Dice exponent $\gamma = 0.7$, and area-tolerance constants are fixed evaluator choices. The score is bounded by $0 \le S \le 100$, with 100 for perfect predictions and 9 for always-empty output. Zero flawed-image credit with one prediction on each clean image yields $S = 9/e \simeq 3.31$. A perfect flawed component with a zero clean component yields $S = 49$, so the components and recall must accompany the composite score. The stored fields `positive_soft` and `clean_soft` correspond to Q_+ and Q_-.

Executable pseudocode. The following sequence completely determines the point score for valid inputs. The supplementary implementation provides the geometry and input validation.

```
1.  Convert annotated region and prediction boxes to
original-image pixels.
2.  Replace each failed or absent prediction record by an
empty list.
3.  Compute containment, tightness, and pair quality for
every pair.
4.  For each intersected annotated region, choose
max(containment, quality, -index).
5.  Group annotated regions by owner; calculate exact union
geometry per group.
6.  Recompute each annotated region quality with shared
union tightness/penalty.
7.  Average owned qualities for prediction credit; unowned
credit is zero.
8.  Compute positive-image F or clean-image exp(-prediction
count).
9.  Aggregate positive/clean components and the fixed
composite score.
```

B.5 Diagnostics and Failure Accounting

Recall at threshold t is the micro-average over annotated regions of $\mathbf{1}[s_k \geq t]$. A complete miss has no intersecting prediction; an effective miss has $s_k < 0.1$. Found-region mean quality is the mean over $s_k \geq 0.1$. Since it conditions on recovered regions, it must be interpreted alongside recall.

Image-level prediction is "flawed" exactly when $N_i > 0$. Image sensitivity, false-negative rate, clean FPR, and overall decision error follow from that decision. Mean clean-image box count averages N_i over the clean subset. Count MAE averages $|N_i - M_i|$ over all images; signed count bias averages $N_i - M_i$. Table 16 reports the absolute value of the signed mean, separately from MAE.

An unowned prediction is a duplicate false positive when it overlaps an owned annotated region with one-to-one pair quality strictly greater than 0.1; otherwise it is a background false positive. A merged prediction owns at least two annotated regions. A split candidate is an annotated region covered by at least two predictions with one-to-one pair quality strictly greater than 0.1. The strict comparison in these auxiliary counts is distinct from the inclusive threshold in region recall.

For family group $h \in \{\text{vis}, \text{text}\}$, collect the corresponding annotated regions on each image. With $K_{i,h} > 0$ regions and mean final quality $\bar{s}_{i,h}$, the family-quality diagnostic is

$$H^h = 100 \frac{\sum_{i:K_{i,h}>0} \sqrt{K_{i,h}}\, \bar{s}_{i,h}}{\sum_{i:K_{i,h}>0} \sqrt{K_{i,h}}}. \tag{18}$$

The stored `artifact` and `text` fields correspond to H^{vis} and H^{text}. Grouping depends only on annotated families, using the same final qualities as the main scorer. These scores contain neither the prediction-precision term nor the clean-image term, so averaging them cannot reconstruct S. Family accuracy is measured only on found regions with a recognized predicted family. Natural-language descriptions are excluded from all scores.

For `status` other than `ok`, missing rows, and absent predictions, $N_i = 0$. Failed flawed images receive $F_i = 0$ and failed clean images receive $F_i = 1$; every benchmark image remains in the denominator.

Maximum score credit from failed clean outputs. Let n_{fail} count all failed image outputs. Setting failed clean images' contributions to zero instead of one, while keeping all images in the denominator, reduces Q_- by at most $d = \min\{n_{\text{fail}}/n^-, Q_-\}$. Holding Q_+ fixed in Eq. (3) gives

$$\Delta S_{\text{fail}} \leq 100\left[(1-\lambda)^2 d + 2\lambda(1-\lambda)\sqrt{Q_+}\left(\sqrt{Q_-} - \sqrt{Q_- - d}\right)\right]. \tag{19}$$

This worst case treats every failed output as clean and removes all of its empty-set credit. With $n^- = 321$, the bounds in S points are 1.448 for Full-Seq (22 failures), 0.390 for Conservative

(7), 0.129 for Trajectory-SFT-R2/Selective (2), 0.064 for Full-Sched (1), and 0.059 for Decision-SFT/Selective (1). The bounds account for four-decimal component rounding and are rounded upward to three decimals. Table 6 keeps the measured scores and reports failure counts separately.

B.6 Bootstrap and Numerical Validation

For each of 2,000 replicates, sample n^+ flawed images and n^- clean images independently with replacement from their respective subsets. Recompute the complete composite score, including the $\sqrt{M_i}$ weights. The 2.5th and 97.5th empirical percentiles form the 95% interval. For sorted replicate values $z_{(0)}, \ldots, z_{(B-1)}$, percentile p uses $h = (B-1)p$ and linear interpolation between indices $\lfloor h \rfloor$ and $\lceil h \rceil$.

All 17 composite scores are consistent with their rounded components: the maximum reconstruction residual is 0.006514, within component/score rounding bounds. The scorer passes 24 synthetic conformance tests covering geometry, ownership, grouping, coordinate conversion, and failure handling.

C Checkpoint Lineage and Training Data

C.1 Supervision and Training Records

Table 4 gives the supervision targets and training configurations of the six fine-tuned checkpoints. Workflow-SFT trains generation-workflow trajectories; the two trajectory critics use different repeat and validation configurations. Decision-SFT continues Workflow-SFT with image-conditioned decision-and-region targets. Balanced+ and Pool+ are described at the scale level: approximately 6,000 labels per class for Balanced+ and more than 20,000 examples for Pool+.

C.2 Workflow-SFT Mixture

Three source batches contain 2,432, 5,055, and 1,347 original conversations; repetition yields 2,824, 5,394, and 4,094 training records. In the first two batches, coordinator records are repeated twice and other roles once; in the third, those factors are four and three. Table 11 gives the role distribution.

Table 11: Research and design account for 57.4% of repeated Workflow-SFT records. Shares use the 12,312-record repeated mixture.

Role	Original	Repeated	Share
Researcher	2,628	3,778	30.69%
Designer	2,559	3,293	26.75%
Intent normalization	1,593	2,121	17.23%
Main coordinator	784	1,674	13.60%
Panel illustrator	879	931	7.56%
Storyboard scripter	182	240	1.95%
Backdrop artist	135	177	1.44%
Shell executor	47	63	0.51%
Storyboard verifier	27	35	0.28%
Total	8,834	12,312	100.00%

Researchers gather or verify evidence; designers plan and execute image generation/editing with self-checks. Other roles handle request normalization, coordination, local generation, execution, and final acceptance. All 27 original storyboard-verifier records have PASS outcomes. Preprocessing retains successful sessions, removes failed/retried/fallback records under the quality rules, validates schemas and message structure, and applies role repetition. Tool calls and image paths are serialized as text. This mixture supervises workflow protocols without explicit defect-region coverage targets.

C.3 Decision-SFT Mixture and Target Granularity

Decision-SFT uses 2,000 positive and 2,000 negative training labels, plus 400 of each for validation. The training and validation image sets are disjoint. Table 12 combines candidate pools, sampling counts, and the positive target-count distribution.

Table 12: Balanced sampling produces 4,000 training and 800 validation records. Negative labels are construction-assigned. The lower panel counts positive images by target-region count.

Source pool	Eligible	Train	Validation
Wikimedia negative	7,077	1,000	200
Pinterest negative	7,311	1,000	200
High-precision non-text positive	117	100	17
Broad positive pool	6,796	1,900	383
Sampled total		4,000	800

Regions/image	1	2	3	4	5	6	7	8	9	10	11
Train	712	723	285	154	66	33	13	9	3	1	1
Validation	152	140	62	20	16	3	5	0	1	0	1

Positive targets. The non-text filter requires confidence at least 0.75, supported verification, major materiality, explicit anomaly evidence, and no hedging. It retains 117 images with 121 edit regions and 139 issues. The broad pool contains 6,802 images predicted flawed; excluding six containing only text-content-fidelity issues leaves 6,796 eligible positives. It includes Qwen-Image variants, U1, and SenseNova-U1.5.

For each edit region, the highest-confidence eligible issue supplies the target region, family, and evidence description. Multiple issues sharing an edit region are compressed into one target. The integer histogram gives 4,368 training and 851 validation target regions, averaging 2.184 and 2.1275 per positive image. All negative targets contain zero regions.

Construction-assigned negative targets. A prescreener had flagged the Wikimedia and Pinterest candidate files as containing defects. Construction retains their image paths, discards the prescreener predictions, and assigns `no_error=true` with an empty list. The eligible pools contain 12,387 and 18,183 prescreener predictions, respectively; these are not human-confirmed defect or mislabel counts. Such negative targets differ from the cross-reviewed benchmark controls.

Training interface. The interface contains only `messages` and `images`, excluding generation prompts, source/model metadata, prescreener decisions as input fields, reasoning targets, and requested-string-fidelity targets. The packed cache contains 1,089 training and 217 validation sequences. Training uses one epoch, 8,192-token packing, and 136 optimizer updates. The supervision supplies image decisions and localized evidence, without region-complete inspection or rejection trajectories.

D Procedure Contracts and Execution Policies

D.1 Shared Records and Execution Policies

The framework separates inventory state, acquired evidence, candidate verification, and final predictions. The inventory kinds shown in Figure 2 are Zone (spatial partition), Entity (object or entity group), Text (text region), and Blind (potentially overlooked area). Review states are unreviewed, verified clear, suspicious, and uncertain. Candidate identities form $\mathcal{V}$, and accepted identities form $\mathcal{A} \subseteq \mathcal{V}$. Table 13 defines the shared records; Table 14 gives execution policies. The evaluated pairings appear in Table 5.

Batch results use `batch_region_id` or candidate IDs rather than positional matching. FullSched caches crops and encodings, overlaps failure handling with inspection, and returns compact

Table 13: Shared logical records and invariants.

Record	Fields / purpose	Invariant
Inventory unit	ID, kind, region, review state, evidence links	Unreviewed is distinct from verified clear
Multi-scale evidence	Overview, context crop, detail crop, transforms	Locations map back to the original image
Candidate	ID, family, box, visible evidence	A proposed finding is not automatically accepted
Verification	Supported / rejected / uncertain; possible box correction	Uncertainty is not affirmative visual evidence
Adjudication	Accepted candidate IDs $\mathcal{A} \subseteq \mathcal{V}$	No new unverified candidate identities
Consolidation	Final groups and source IDs	Preserve or explicitly exclude candidate provenance
Failure handling	Invalid IDs, missing results, failed required work	A failed call is not evidence of cleanliness

Table 14: Operational comparison of multi-stage procedures. Full review routes every required inventory unit to local inspection. Selective changes this routing through triage; Full-Seq and Full-Sched differ in scheduling.

Aspect	Full-Seq	Full-Sched	Selective
Local-review policy	Full required inventory	Full required inventory	Triage-selected units
Inventory execution	Sequential full-review path	Independent inventories overlap	Adds review-required and confidence fields
Local inspection	Region-wise multi-scale review	Text batches of 2; artifact batches of 4	Dense-path batching; skipped units tracked
Verification	Candidate-wise verification	Up to 2 candidates per verification batch	Verified-local / peer-scout fast paths; verification batches otherwise
Image preparation	Context/detail views	Cached, asynchronously prepared views	Inherits cached multi-scale views
Consolidation	Online grouping and provenance check	Compact source-ID grouping; text backfill	Component repair / fallback on invalid grouping

consolidation decisions. Existing evidence text is restored through `source_finding_id` rather than regenerated.

Selective uses `review_required` and `triage_confidence` to route suspicious, uncertain, small, complex, or occluded units to detail review. High-confidence clear units may bypass it. For text, text-crosscheck, or entity groups with at least 12 regions, the dense path uses batches of at least four. Returns violating identity or completeness requirements are rejected; failure handling can split a failed three-item batch into two plus one. The structure scout checks geometry and relations; the rendering scout checks rendering and text, providing discovery paths alongside local multi-scale review.

Selective uses local-verification and peer-scout fast paths with deterministic adjudication; remaining candidates undergo verification. Figure 2 depicts these functional stages. A candidate supported by an eligible fast path does not require another independent verification call.

D.2 Conservative Comparator

Conservative denotes the v5 multi-stage implementation (`u1_critic_v5_20260807`) run with Workflow-SFT weights and the common region-set output contract. It is a separate procedure from Full-Seq, Full-Sched, and Selective. Relative to Direct, it returns 0.430 rather than 0.832 predictions per clean image, with clean FPR of 25.2% versus 35.5% and R@0.1 of 23.0% versus 24.4%.

E Complete Supplementary Results

Tables 15 and 16 supplement the 17 configurations in Table 6. Component scores and intervals retain additional precision; image false-negative rate is the complement of sensitivity.

Table 15: Composite-score intervals and components distinguish recovery from clean-image credit. H^{vis} and H^{text} group annotated families as defined in Appendix B; they exclude prediction precision and the clean component.

Model / procedure	S	95% CI for S	H^{vis}	H^{text}	$100Q_+$	$100Q_-$
Qwen3.6-27B	23.373	[21.780, 24.907]	13.189	42.058	21.15	28.99
Qwen3.8-27B	23.920	[22.379, 25.527]	11.243	31.716	18.77	38.37
Qwen3.7-Plus	19.665	[18.116, 21.207]	14.695	26.249	15.88	30.06
GPT-5.6-sol	20.240	[18.643, 21.742]	3.558	9.264	5.75	88.39
Gemini-3.5-Flash	16.269	[15.031, 17.571]	5.121	5.170	5.39	64.47
Workflow-SFT	28.508	[26.671, 30.294]	9.709	35.222	16.16	70.87
Trajectory-SFT-R1	24.535	[22.806, 26.243]	14.815	43.953	21.57	32.19
Trajectory-SFT-R2	25.406	[23.685, 27.131]	14.648	42.931	21.49	35.81
Decision-SFT	30.419	[28.598, 32.190]	10.120	26.929	14.21	91.92
Decision-SFT-Balanced+	22.387	[20.793, 24.054]	6.460	4.650	6.42	97.24
Decision-SFT-Pool+	19.173	[17.689, 20.581]	3.925	2.241	4.01	98.42
Workflow-SFT / Full-Seq	31.549	[29.810, 33.317]	15.789	43.374	20.71	65.66
Workflow-SFT / Conservative	28.738	[26.994, 30.489]	6.278	38.412	14.35	81.56
Workflow-SFT / Full-Sched	31.895	[30.066, 33.725]	14.799	43.929	20.17	69.65
Workflow-SFT / Selective	29.406	[27.727, 31.190]	14.169	37.550	18.64	64.04
Trajectory-SFT-R2 / Selective	29.059	[27.254, 30.821]	14.056	41.474	18.76	61.83
Decision-SFT / Selective	31.403	[29.718, 33.197]	12.724	36.028	17.29	80.60

Table 16: Final-output diagnostics expose decision, localization, and count errors. Decision error and no-overlap are percentages; clean predictions are averaged per clean image. Quality conditions on $s_k \geq 0.1$. MAE and |bias| concern counts; BG FP counts all unowned background predictions and BG FP$^+$ those on flawed images.

Model / procedure	Dec. error	Clean pred.	No overlap	Quality	MAE	|bias|	BG FP	BG FP$^+$
Qwen3.6-27B	29.84	2.2056	50.18	0.6843	2.1444	0.8373	1939	1231
Qwen3.8-27B	28.97	1.1433	59.76	0.6678	1.4755	0.3686	870	503
Qwen3.7-Plus	26.08	2.0717	48.87	0.6373	1.8797	0.9249	1983	1318
GPT-5.6-sol	39.17	0.2555	78.42	0.6402	1.6092	0.9798	544	462
Gemini-3.5-Flash	19.35	0.7477	73.47	0.4198	1.3763	0.3561	1153	913
Workflow-SFT	30.51	0.8318	60.11	0.6735	1.6853	0.1396	1089	822
Trajectory-SFT-R1	27.24	2.0093	43.97	0.6687	1.8518	0.6660	1671	1026
Trajectory-SFT-R2	27.05	1.6854	44.93	0.6686	1.7045	0.4418	1466	925
Decision-SFT	30.32	0.1994	62.78	0.6648	1.5236	0.5881	706	642
Decision-SFT-Balanced+	49.18	0.0436	85.73	0.6390	1.6939	1.4880	186	172
Decision-SFT-Pool+	56.50	0.0249	91.98	0.6300	1.7603	1.7141	66	58
Workflow-SFT / Full-Seq	24.93	0.8629	45.44	0.6536	1.5140	0.0356	1059	782
Workflow-SFT / Conservative	38.02	0.4299	63.34	0.6469	1.5274	0.7517	568	430
Workflow-SFT / Full-Sched	23.48	0.8069	46.75	0.6678	1.5544	0.1357	1173	914
Workflow-SFT / Selective	23.20	0.8692	46.95	0.6403	1.6622	0.2685	1326	1047
Trajectory-SFT-R2 / Selective	25.89	1.1121	45.34	0.6428	1.8537	0.4870	1531	1174
Decision-SFT / Selective	23.48	0.4891	50.38	0.6449	1.6131	0.0269	1088	931

All clean-image predictions are background false positives. Their total is $\mathrm{round}(321\bar{N}_-)$. Subtracting this total from BG FP gives BG FP$^+$. Because the published clean mean is rounded to four decimals, its count uncertainty is at most $321 \times 0.00005 < 0.017$; the compatible integer is unique for every configuration.

Family-group diagnostic. Text-group quality exceeds visual-defect quality in 15 of 17 configurations; Full-Sched has $H^{\text{text}} = 43.9$ versus $H^{\text{vis}} = 14.8$. These are pooled family summaries across different source mixtures. Count MAE summarizes absolute count error, while the absolute signed mean bias summarizes the magnitude of net over- or underprediction.

F Representative Benchmark Examples

Figure 3 shows five representative flawed images with human-annotated regions. The colored outlines follow the five families in Table 8: Rendering, Structural, Topology, Relational-physical, and Text integrity. These examples illustrate three properties of the benchmark. First, one image can contain multiple independently localizable defects; the five panels contain 6, 5, 4, 7, and 5 annotated regions, respectively, in reading order of top-left, bottom-left, top-right, middle-right, and bottom-right. Second, annotation targets vary widely in scale, from small local text or rendering problems to larger structural or relational failures. Third, a single image may mix several families, so inspection requires localized evidence rather than only an image-level judgment.

The overlays indicate human annotation scope only and should not be interpreted as model outputs. Coupled symptoms may share one region when they form a single localized defect, whereas spatially separable evidence is marked independently. The figure therefore complements the aggregate statistics in Appendix A by showing how the benchmark's region sets appear in practice.



Figure 3: Representative INSPECTBENCH examples with human-annotated defect regions. Colors correspond to the five defect families listed in Table 8. The five panels contain 6, 5, 4, 7, and 5 annotated regions, respectively, in reading order of top-left, bottom-left, top-right, middle-right, and bottom-right.

G Extended Related Benchmarks and Evaluation Context

Public-4 source tasks. Table 17 summarizes Public-4 tasks before common image-only annotation (Section 3).

Table 17: Public-4 source tasks before common image-only annotation and cross-review.

Source	Original task	Annotated evidence / target
ArtiBench (Park et al., 2026)	Artifact assessment	Presence labels, regions, explanations
SalArt-VQA (Sun et al., 2026)	Salient-artifact VQA	Selected artifact, spatial choices, evidence
X-AIGD (Xiao et al., 2026)	Synthetic-image detection	Categorized pixel masks
TextPecker (Zhu et al., 2026)	Character-structure assessment	Character-level anomalies

Composition, text, and physical consistency. DSG organizes dependent prompt questions (Cho et al., 2024); DPG-Bench tests dense prompt requirements (Hu et al., 2024). ImagenWorld provides localized human error annotations (Mahdizadeh Sani et al., 2026). TextAtlasEval tests long, structured text generation (Wang et al., 2025a); OCRGenBench covers text generation and editing (Zhang et al., 2025). SciFlow-Bench uses canonical graph structure (Zhang et al., 2026b). PICABench grounds physical evidence in regions (Pu et al., 2025); PhyEditBench uses video-grounded edits (Guo et al., 2026); CPI-Bench includes multi-image and reasoning-intensive editing (Zhou et al., 2026). Their references establish requirements beyond the image alone.

Quality and grounded judgments. Q-Bench evaluates low-level vision; Q-Instruct supplies instruction data (Wu et al., 2024a;b). Q-Align learns quality levels, and DepictQA gives perceptual explanations (Wu et al., 2024c; You et al., 2024). Grounding-IQA evaluates localized descriptions, questions, and grounding precision (Chen et al., 2026), complementing region-set recovery.

Preference versus region inspection. CLIPScore measures caption compatibility and HPS v2 models human preferences (Hessel et al., 2021; Wu et al., 2023); both complement quality judgments rather than enumerate visible defects. Model Cards motivate disclosing the operating scope of critic baselines (Mitchell et al., 2019).

Geometric evaluation and diagnostic decomposition. Generalized IoU studies localization geometry (Rezatofighi et al., 2019), while TIDE motivates interpretable error categories (Bolya et al., 2020). Here ownership and union recalculation determine region credit, while the outcome partition separates absent overlap from insufficient credit (Appendix B).

Normal controls and false visual claims. MVTec AD provides normal and anomalous industrial images; LOCO AD tests structural and logical anomalies (Bergmann et al., 2019; 2022). Hallusion-Bench diagnoses entangled hallucination and illusion (Guan et al., 2024). Our clean FPR measures nonempty defect predictions (Section 3.4).

Multi-stage inspection as a comparison axis. V* and Visual CoT use visual search or focus for reasoning (Wu & Xie, 2024; Shao et al., 2024). ArtiAgent synthesizes artifact-training data through perception, synthesis, and curation (Park et al., 2026); ArtifactLens combines modular detection, in-context learning, and instruction optimization (Burgess et al., 2026); Agentic Retoucher stages perception, reasoning, and repair (Shen et al., 2026). These motivate the inspection and supervision axes in Table 5.