SCIFIGQUAL-BENCH: SCIENTIFIC FIGURE QUALITY ASSESSMENT WITH MANUSCRIPT-LINKED CONTEXT

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific figure quality is bound to the manuscript: a crop can be visually clear and still contradict its caption or the paragraph that cites it. Most image quality assessment and chart-understanding methods target a detached visual surface or a question-answering score, rather than alignment among the figure, its caption, and the citing text. A single end-to-end prompt that sees the image and the text together still mixes what is visible with what the caption and the citing paragraph claim, so fluent wording can raise a visual score and missing text can be treated as low quality. To address this gap, we introduce SciFigQual-Bench, a manuscript-linked benchmark for published CS-conference figures that binds each figure to its caption and to index-resolved citing paragraphs, and scores visual clarity, layout, caption consistency, context consistency, and misleading risk, leaving a dimension unevaluated when its evidence is absent. SFQ-Agent reads the image and the text in separate calls and records modality-specific evidence. A cross-modal judge scores caption and citing-text alignment from those reports, and a deterministic runner keeps the visual scores, sets misleading risk by a fixed rule, and applies the written caps, so each dimension follows the rubric instead of a single end-to-end prompt. Experiments comparing this staged judge with single-pass and OCR-sidecar protocols find the closest point-estimate fit to mean human ratings under staged judging, while the gap between protocols remains small. Caption consistency remains the main gap, and agreement with the rater mean answers a different question from agreement among human raters.

1 INTRODUCTION

Natural photographs are judged mainly by human visual perception: in image quality assessment (IQA), a landscape is scored on color, sharpness, and aesthetics, with no content-level right-or-wrong standard (Li et al., 2025), and can often be evaluated from pixels alone (Chen et al., 2024). Scientific figures, by contrast, are bound to academic text (Tufte, 2001; Cleveland & McGill, 1984): truncated axes and missing baselines are intrinsic visual-design defects readable from the canvas, whereas contradictions with captions or citing prose require manuscript context beyond pixel-only inspection.

In peer review, figure quality is therefore inherently tri-modal: it depends on what is visible in the image, what the caption claims, and what the manuscript asserts in citing paragraphs. Existing natural-IQA and AIGC-alignment paradigms operate on a single visual modality and decouple the image from both its caption and the citing manuscript. As illustrated in Figure 1, four mainstream paradigms highlight this gap: Natural IQA (Mittal et al., 2012; 2013; Prashnani et al., 2018; Wang et al., 2023) estimates perceptual quality from low-level statistics. AIGC alignment (Radford et al., 2021; Xu et al., 2023; Wu et al., 2023; Hu et al., 2023; Kirstain et al., 2023) measures text and image similarity for generated content. Schematic understanding (Masry et al., 2022; Methani et al., 2020; Kafle et al., 2018; Liu et al., 2023; Masry et al., 2023) frames classification or visual question answering without tying model judgments back to the citing paragraphs of the source PDF. Figure reasoning QA (Siegel et al., 2016; Mathew et al., 2021) tests whether a model can read a figure but does not offer a rubric for holistic quality or misleading risk. In each case, evaluation proceeds on a detached visual input.

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

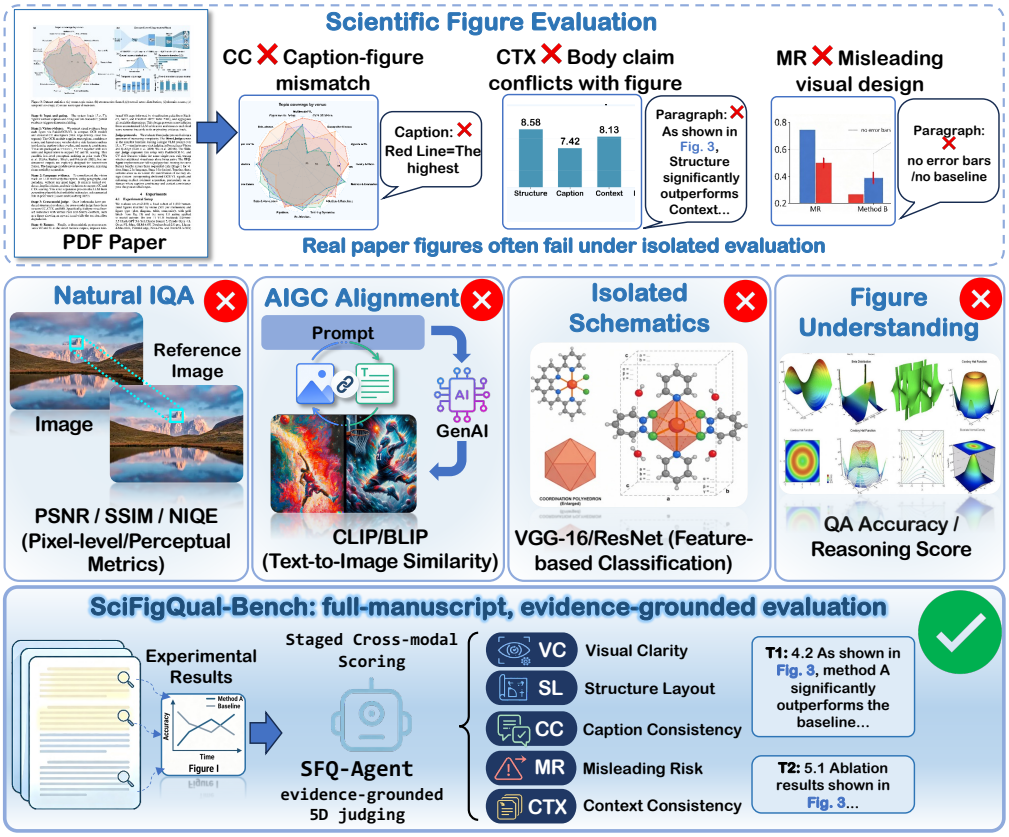


Figure 1: From isolated-figure evaluation to manuscript-linked context binding. Prior paradigms evaluate figures without manuscript evidence, whereas SciFigQual-Bench grounds assessment in figure-relevant citing text and scores five separately gated dimensions on a unified 1 to 10 scale.

In this paper we propose SciFigQual-Bench for human-published scientific figures with manuscript-linked context, binding each figure to its caption and the citing-paragraph set $\mathcal{T}$ resolved from the source PDF when in-text figure references exist (otherwise $\mathcal{T}=\emptyset$ and CTX is L1-gated). Judge prompts receive at most two bounded citing snippets rather than the full manuscript, so our claim is contextual linkage, not long-context whole-paper reasoning. Instead of a single end-to-end vision and language model (VLM) prompt, we collect multi-modal evidence from image, caption, and citing text, fuse it, and output traceable scores on Visual Clarity (VC), Structure & Layout (SL), Caption Consistency (CC), Context Consistency (CTX), and Misleading Risk (MR). Unlike faithfulness benchmarks for AI-generated scientific images (Ye et al., 2026; Li et al., 2024), the task is retrospective assessment of published figures and is complementary to generation-faithfulness evaluation. Our contributions can be summarized as follows:

- We release a manuscript-linked benchmark. Published-figure quality is a judgment over the figure, its caption, and index-resolved citing paragraphs, with five dimensions and L1 gating. SciFigQual-Bench provides 7,609 curated figures from 1,144 papers (6,308 human-rated) in top-tier CS conferences. CTX is scored only when a citing paragraph is present.
- We build a traceable judge. SFQ-Agent collects visual evidence and textual evidence in separate calls and combines them with a deterministic runner, so VC, SL, CC, CTX, and MR each follow the written rubric. A single-pass Direct prompt and an OCR sidecar are run on the same figures, which makes the effect of staging measurable.
- We measure mean fit against a rater ceiling. The frozen *eval1200* split has 1,200 figures, dimension-first human gold, and the same L1 mask at inference. Model Within-1 is taken against that mean, and pairwise human agreement on the same split is reported as the rater ceiling.

Table 1: Capability and scale comparison against related benchmarks. Symbols: ✓=fully supported. △=partial. ✗=not targeted. Ours reports 7,609 released figures (6,308 rated) with per-figure captions and bound citing paragraphs $\mathcal{T}$ when present (≈7.6k released text attachments). The ≈355k citing-paragraph records and 62,694 PDFs are crawl-stage construction totals, not the released inventory. Other columns follow each paper's reported test or release size.

Comparison dimension	SIQA	SCIEval	VisJudge	SPUR	GENFIG1	Ours
Real figures from published CS papers	✗	✗	△	✗	✗	✓
Source PDF + section binding	✗	✗	✗	△	△	✓
Full-text citing paragraphs (index-driven)	✗	✗	✗	△	✗	✓
Evaluation dimensions (multi-dim. rubric)	✓	✓	✓	△	✓	✓
End-to-end Acquire→Annotate pipeline	✗	△	✗	✗	△	✓
Auditable scoring rationale	△	△	△	✗	△	✓
Multi-venue × multi-year (CS top-tier)	△	✗	△	✗	△	✓
# Images (released)	2,240	6,000	3,090	1,084	584	**7,609**
# Text items (released)	≈2.2k	≈6.0k	≈3.1k	≈4.3k	≈584	**≈7.6k**
# Papers (crawl scope)	n/a	n/a	n/a	≈5k	≈584	62,694

2 Related Work

Existing work that can be applied to scientific figure quality assessment and multimodal figure and text evaluation falls into four categories: perceptual IQA, chart comprehension, text-to-figure generation and multimodal corpora, and LLM-based automated evaluation.

Perceptual IQA and Isolated Image Quality. Classical no-reference IQA (BRISQUE, NIQE) and learned perceptual metrics (PieAPP, CLIP-IQA) estimate quality from low-level statistics or deep features (Mittal et al., 2012; 2013). SIQA defines a four-dimensional rubric for schematic scientific figures and trains learned regressors on expert ratings, yet its instances are not drawn from published CS PDFs and carry no index-resolved citing paragraphs (Li et al., 2026). VisJudge-Bench employs MLLM judges for visualization aesthetics on chart and diagram crops (Xie et al., 2026), but omits source-PDF binding and manuscript-level CC/CTX. These approaches remain visual-only on isolated crops, and we likewise do not train a specialist quality regressor: the experiments test prompted protocols on the same frozen split.

Chart and Scientific Figure Understanding. ChartQA and SciFIBench benchmark chart reasoning and multiple-choice scientific-figure interpretation (Masry et al., 2022; Roberts et al., 2024). SPUR extends to biomedical experimental images with perception, understanding, and reasoning QA and partial figure and text linkage (Ding et al., 2026), yet targets VQA accuracy rather than a calibrated multi-dimensional quality rubric and covers neither CS top-tier venues nor index-driven citing paragraphs at our scale. These resources analyze scientific images but do not target caption and figure match, manuscript consistency, or misleading risk on official CS-conference figures.

Generation Faithfulness. SCIEval trains and benchmarks CLIP/LMM modules on multi-dimensional faithfulness of generated scientific images, covering relevance, accuracy, and interpretability in text-to-image and captioning settings (Ye et al., 2026). GENFIG1 asks VLMs whether automatically produced visual summaries faithfully reflect paper content (Guan et al., 2026). At only hundreds of instances, it targets quality control of *newly synthesized* academic illustrations rather than retrospective assessment of human-authored conference figures, so extraction and captioning alone do not certify figure quality in peer-review terms.

LMM-as-Judge. Rubric-conditioned VLM evaluators (Prometheus-Vision, VIEScore) enable fine-grained multimodal scoring (Lee et al., 2024; Ku et al., 2024). Evidence-grounded NLP cautions against plausible but unfaithful rationales (Jacovi & Goldberg, 2020). Monolithic judges conflate visual perception with textual verification and can hallucinate evidence when scoring CC, CTX, and MR jointly.

SciFigQual-Bench links each figure to its citing paragraphs resolved from the PDF, gathers modality-specific evidence stepwise, and fuses them only after these stages, targeting real published figures rather than isolated crops, synthetic substitutes, or multiple-choice interpretation. Table 1 compares the benchmark with the five closest 2024 to 2026 efforts.

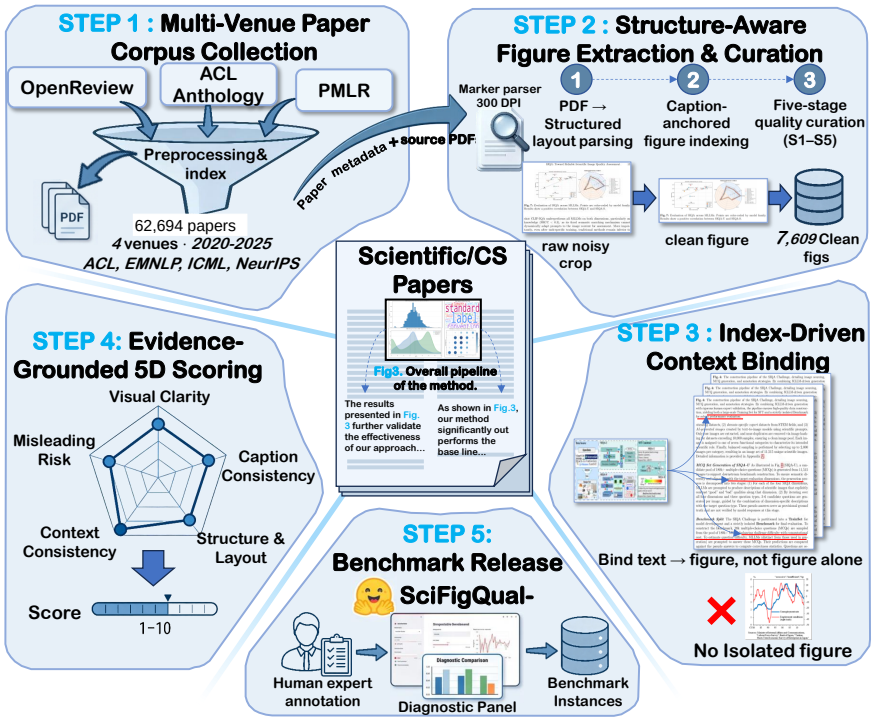


Figure 2: SciFigQual-Bench construction pipeline: corpus collection, structure-aware extraction, context binding, five-dimensional rubric, and human validation.

3 SciFigQual-Bench

This section defines benchmark instances and the five-dimensional rubric, then describes data construction and the SFQ-Agent scoring framework.

3.1 Task Formulation and Rubric

A benchmark instance is a tuple $(I, c, \mathcal{T}, m)$ where I is the figure crop, c the caption (possibly empty), $\mathcal{T}$ the citing paragraphs (possibly empty), and m lightweight metadata (venue, year, figure index, section tags). Instances are always anchored to a source PDF so that $\mathcal{T}$ is resolved by figure index rather than layout heuristics alone. Human annotators and SFQ-Agent assign scores on five separately gated dimensions, each on $[1, 10]$: Visual Clarity (VC), the legibility of text, marks, and encoding at publication scale. Structure & Layout (SL), composition, panel organization, and chartjunk (Tufte, 2001). Caption Consistency (CC), alignment between c and visible content, including subpanels and trends. Context Consistency (CTX), alignment between evidentiary claims in $\mathcal{T}$ and what I supports. Misleading Risk (MR), the likelihood of reader misinterpretation (higher score = lower risk), including truncated axes or undisclosed baselines (Rückert et al., 2025). Overall quality is the mean over available dimensions and is used for protocol ranking when captions or citing text are missing. Let $\mathcal{D} = \{\mathrm{VC}, \mathrm{SL}, \mathrm{CC}, \mathrm{CTX}, \mathrm{MR}\}$ and $g_i(d) \in \{0, 1\}$ indicate whether dimension d is evaluable for instance i under L1 gating, with human and model outputs as vectors $\mathbf{s}_i, \hat{\mathbf{s}}_i \in [1, 10]^{|\mathcal{D}|}$. The gated overall label is

$$y_i = \frac{1}{\sum_{d \in \mathcal{D}} g_i(d)} \sum_{d \in \mathcal{D}} g_i(d)\, s_i(d), \tag{1}$$

with y_i undefined only when both c and $\mathcal{T}$ are absent (the instance is excluded). If $c{=}\emptyset$, CC is marked null ($g_i(\mathrm{CC}){=}0$). If $\mathcal{T}{=}\emptyset$, CTX is null. This L1 evidence gating prevents treating missing metadata as low quality and mirrors annotation practice in SciCap+ (Yang et al., 2024), while MR may still be scored from visual hazards when text is partial.

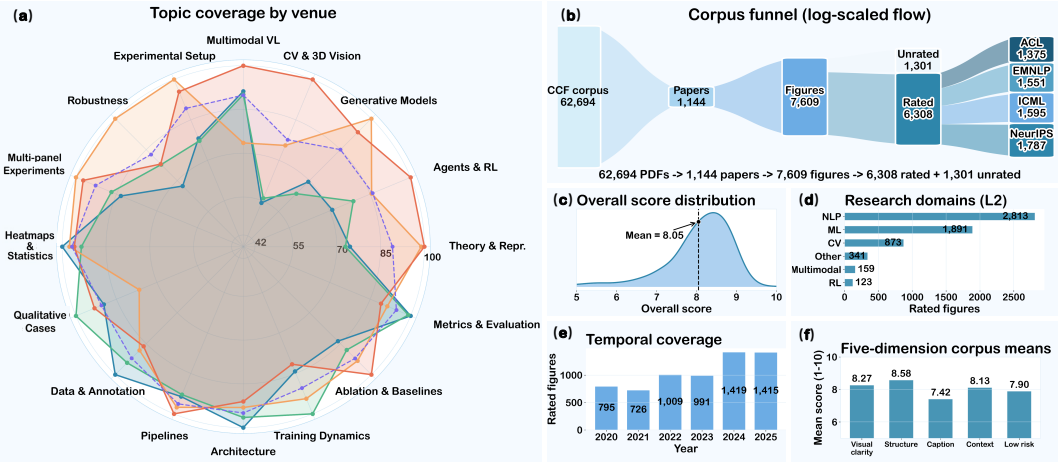


Figure 3: Corpus overview of SciFigQual-Bench. From 62,694 PDFs we retain 1,144 papers and 7,609 figures (6,308 human-rated) across ACL/EMNLP/ICML/NeurIPS (2020 to 2025). Overall scores concentrate near 8.05. Caption consistency is the weakest axis.

3.2 Benchmark Construction

Figure 2 summarizes the pipeline over CS papers from ACL, EMNLP, ICML, and NeurIPS (2020 to 2025). We formalize the released benchmark as $\mathcal{B} = \{(I_j, c_j, \mathcal{T}_j, m_j)\}_{j=1}^{N}$ with N=7,609 figures from P=1,144 qualified papers, of which N_{rat}=6,308 carry human five-dimensional scores, while the remaining 1,301 figures retain image, caption, and context bindings for protocol development and future labeling but are excluded from gold-based evaluation on *eval1200*. Each module below is deterministic given source PDFs and fixed preprocessing seeds.

Step 1: Corpus Acquisition. We crawl PDFs and bibliographic metadata from OpenReview, ACL Anthology, and PMLR, tagging each file with $(venue, year, paper_id)$. PDFs are linearized, publisher watermarks are stripped where possible, and exact duplicates are removed by content fingerprint $h(p) = \text{SHA256}(\text{bytes}(p))$, after which records outside 2020 to 2025 or missing parseable metadata are discarded. The surviving raw corpus is

$$\mathcal{C}_0 = \{p_k\}_{k=1}^{K_0}, \quad K_0 = 62{,}694, \tag{2}$$

larger in crawl scope than DocFigure and SciFIBench sources (Jobin et al., 2019; Roberts et al., 2024; Li et al., 2024), but used here for quality evaluation rather than MCQ or type classification.

Step 2: Structure-Aware Figure Extraction. We rasterize each PDF at 300 DPI and parse its layout with Marker, obtaining text blocks, figure bounding boxes, and caption candidates, which are ranked by spatial overlap, vertical distance, and a figure-caption cue (Song et al., 2025). A five-stage curation pipeline then removes duplicates and non-figures, filters anomalous crops, recovers clipped panels, flags borderline cases, and discards irrecoverable layouts (Appendix A.3).

Step 3: Index-Driven Context Binding. We detect figure references in the body with PyMuPDF by matching index patterns, collect the hosting paragraph for each match, and merge multi-sentence windows within the same section, yielding per-figure context sets $\mathcal{T}_k = \bigcup_{\tau \in \text{match}(k)} u(\tau)$, each capped at $L_{\max}$ tokens. Index matching over the raw crawl yields approximately 355,000 citing-paragraph records. These are index hits rather than abstract or introduction heuristics (Wang et al., 2026; Li et al., 2024). After curation, each of the 7,609 released figures is paired with its caption and the resolved $\mathcal{T}$ (possibly empty). Scoring prompts use at most two snippets (Appendix A.4). Paragraphs lacking explicit figure-specific claims are down-weighted, and comparative or trend language is preferred for subsequent CTX annotation.

Step 4: Five-Dimensional Human Annotation. Trained human annotators were calibrated on a 50-figure pilot using the rubric in Appendix A.1, after which each instance received scores under Eq. 1 with L1 gating, so that CC and CTX were hidden whenever the necessary evidence was absent. Following faithful-NLP guidelines (Jacovi & Goldberg, 2020), annotators had to justify every

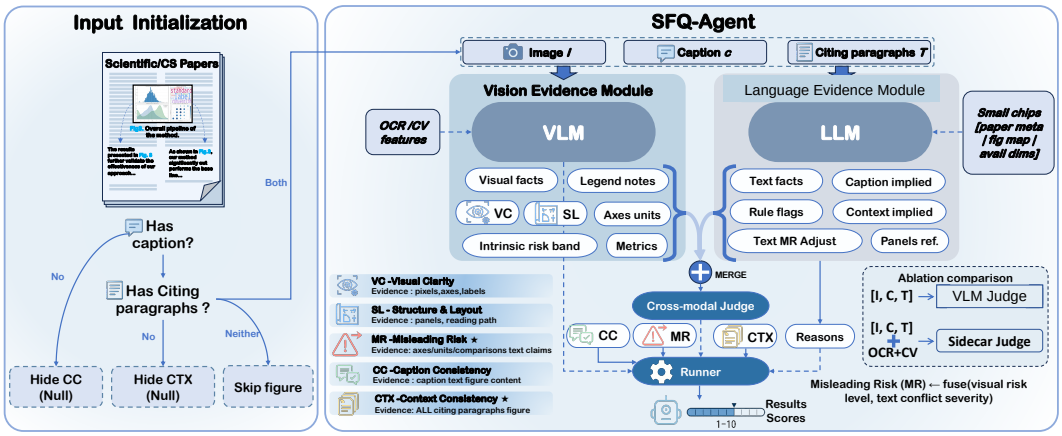


Figure 4: SFQ-Agent scoring pipeline: L1 gating, parallel vision and language evidence modules, cross-modal judge, and deterministic Runner aggregation. Direct and Sidecar judges serve as ablations.

score with specific visible marks, caption phrases, or citing sentences, and any rationale lacking evidence was returned for revision. We performed dual annotation on an ACL 2025 holdout, with disagreements exceeding two points triggering adjudication. The released gold for *eval1200* uses at least three raters per figure, aggregated dimension-first (Appendix D.3). On complete ratings this coincides with the mean of per-rater overalls, and the two can diverge only under uneven L1 null masks across raters. The final set comprises 6,308 rated figures.

Step 5: Release and Evaluation Packaging. We package all rated records, context indices $\mathcal{T}$, and model-score outputs in a unified JSONL schema, together with a frozen evaluation split, *eval1200*, of 1,200 paper-aware instances mixed by figure type. Gold labels were frozen before model calls. The split and its gold are released for reproducibility after protocol freeze and should not be treated as a permanently hidden leaderboard test. Three judge protocols (Direct, Sidecar, and SFQ-Agent) share the same input fields $(I, c, \mathcal{T}, m)$ and the same five-dimensional schema, and differ only in evidence staging and call budget.

3.3 Dataset Statistics

Figure 3 summarizes scale, diversity, and quality relative to prior scientific-figure benchmarks. Radar coverage in (a) spans 18 research categories across four venues, broader than chart-only sets such as CharXiv (Wang et al., 2024). The funnel in (b) yields P=1,144 papers and N=7,609 figures from K_0=62,694 PDFs (Eq. 2), of which 6,308 are human-rated, exceeding SIQA-U (2,240) and SPUR (1,084) in comparable settings (Li et al., 2026; Ding et al., 2026). Paper retention is $P/K_0 \approx 1.8\%$, and figure yield after curation is $N/K_0 \approx 0.12$ figures per crawled PDF. Overall scores in (c) concentrate near 8.05 because published figures are mostly high quality, so ranking is harder than coarse Within-1. Rated figures in (d) are dominated by NLP (2,813), ML (1,891), and CV (873), with the remainder spanning other CS areas, while temporal coverage in (e) peaks in 2024 to 2025 with over 1,400 rated figures per year. Mean dimension scores in (f) show CC (7.42) as the weakest axis versus SL (8.58): caption and figure mismatches, rather than blur, are the typical defects in this published-only pool.

3.4 SFQ-Agent

The statistics above identify caption and context consistency as the main bottleneck, which a single end-to-end VLM struggles to score. SFQ-Agent (Figure 4) implements evidence-grounded judging as an auditable alternative to monolithic LMM-as-judge pipelines (Zheng et al., 2023; Lee et al., 2024; Ku et al., 2024).

Step 0: Input and Gating. The system loads $(I, c, \mathcal{T})$, discards figures that lack both a caption and citing text, and hides dimensions when evidence is only partial. **Step 1: Vision Evidence.** We

extract visual evidence via PaddleOCR-VL and classical CV features (blur, edge density, color histogram). The OCR/CV module supplies text regions, confidence, and layout cues for VC/SL, packaged with axis/legend notes as `visual_facts`. Precomputed caption and OCR token/numeric overlap ratios are side features for later CC fusion, not substitutes for pixel-level VC/SL. The language module never accesses pixels. **Step 2: Language Evidence.** An LLM reads only the caption, citing paragraphs, and metadata, with no pixel input, and extracts textual evidence, implied claims, and text-side severity/caps (omissions, bare references) to support later CC/CTX/MR fusion. This separation prevents plausible but unfaithful rationales (Jacovi & Goldberg, 2020). **Step 3: Cross-Modal Judge.** The judge fuses the two evidence reports to score CC/CTX and to write an MR rationale, aligning visual hazard indicators with textual cues and detecting conflicts (e.g., an upward trend in the figure versus "degradation" in the text). VC and SL are copied from Stage 1 rather than re-estimated from text summaries, so fluent language cannot re-inflate visual scores. **Step 4: Runner.** A deterministic post-processor locks VC and SL to the vision-module outputs, replaces the MR number by the closed-form fusion of $[\ell, h]$ and σ (Appendix C.3), applies CC/CTX caps, and aggregates under L1 gating, which prevents score inflation from unconstrained LLM arbitration and keeps each final score traceable to its originating evidence track. Frozen prompt contracts and the MR fusion mapping are reported in Appendix C.

Judge Protocols. We evaluate three protocols of increasing complexity. The Direct Judge issues a single VLM prompt over $(I, c, \mathcal{T})$, similar to zero-shot judging in Prometheus-Vision and Q-Align (Lee et al., 2024; Wu et al., 2024). The Sidecar Judge adds PaddleOCR-VL and CV side features within the same single-pass call. SFQ-Agent reuses the same feature bundle across three sequential calls (vision, language, fusion). The variants isolate OCR-VL side signals versus explicit evidence staging, especially on CC and CTX, where monolithic judges most often conflate perception with text verification.

4 Experiments and Results

4.1 Experimental Setup

We evaluate on *eval1200*: 1,200 figures from 254 papers, sampled paper-wise (ACL 266, EMNLP 294, ICML 294, NeurIPS 346) and mixed by figure type (plot, diagram, table, composite). Gold follows Eq. 1, and model outputs use the same L1 mask (Appendix B.1). CTX is scored on the 807 figures with citing text. Three of them keep only one CTX rating after per-rater nulls, so Table 3 uses the remaining n=804. Eleven backends run under Direct and, where available, Sidecar. SFQ-Agent adds Gemini-3.1-Pro and matches vision with language except for F5 (Qwen-VL-Max vision with Qwen-Plus language). The 29 rows share $(I, c, \mathcal{T}, m)$ and do not fill an 11×3 grid, because Sidecar omits the Claude backends (Appendix B.2). Direct is one VLM call that returns every evaluable dimension together. Sidecar keeps that single call and adds PaddleOCR-VL and CV features, testing whether extra visual cues help without changing the decision. SFQ-Agent uses three calls (Section 3.4): vision locks VC/SL, language reads caption and citing text without pixels, and fusion scores CC, CTX, and MR, after which the Runner locks the visual scores and applies the caps. The comparison therefore separates denser OCR observation from explicit modality separation (Appendix C). Metrics on the full split, unless a table says otherwise, are MAE, Within-1 (within ±1), Spearman correlation, and signed bias (positive = lenient) versus mean gold. CC, CTX, and MR additionally report MAE on L1-evaluable items only. A constant predictor of the split mean 8.13 is the floor (MAE 0.565, W-1 87.9%).

4.2 Protocol Ablation Results

Table 2 consolidates the main evaluation on *eval1200*: 29 protocol and backend configurations spanning Direct, Sidecar, and SFQ-Agent. Rows share $(I, c, \mathcal{T})$ and L1 gating, but Agent uses three calls per figure and Sidecar is missing Claude backends, so rows are comparable on inputs rather than on compute.

Direct Baseline. Claude-Opus-4.8 leads Direct protocols (MAE 0.443, W-1 92.4%, SRCC 0.582), closely followed by GPT-5.6-Sol. Qwen-VL-Max is reasonably calibrated (W-1 88.8%) but ranks poorly (SRCC 0.416), driven by high CC/CTX errors ($\mathrm{MAE}_{\mathrm{CC}} \approx 0.99$, $\mathrm{MAE}_{\mathrm{CTX}} \approx 1.04$), while Doubao-Seed-2.0-pro shows strong negative bias (-0.120), indicating systematic harshness, and

Table 2: Protocol and backend ablation on *eval1200* (n=1200) versus *mean* gold. ↑ better. ↓ worse. W-1 is not pairwise human agreement. Green: best per protocol (lowest MAE). Sidecar "+OCR" uses PaddleOCR-VL.

Backend	MAE↓	W-1↑	SRCC↑	Bias	CC↓	CTX↓	MR↓
Direct (1 call)							
Gemini-3.5-Flash	0.464	91.0%	0.532	0.032	0.942	0.972	0.724
GPT-5.6-Sol	0.448	92.0%	0.565	0.022	0.922	0.928	0.708
Claude-Sonnet-5	0.478	90.2%	0.518	0.038	0.968	0.862	0.732
Qwen-VL-Max	0.520	88.8%	0.416	0.065	0.985	1.044	0.758
GLM-4.6V	0.662	81.0%	0.438	0.102	1.082	1.128	0.812
Doubao-Seed-2.0-pro	0.609	83.2%	0.570	-0.120	0.878	0.918	0.752
Llama-4-Maverick	0.486	89.8%	0.512	0.035	0.958	0.948	0.728
Pixtral-Large	0.502	89.2%	0.498	0.042	0.972	0.968	0.742
Nova-Pro	0.471	90.6%	0.528	0.028	0.938	0.912	0.716
Claude-Opus-4.8	**0.443**	**92.4%**	**0.582**	**0.015**	**0.908**	**0.888**	**0.692**
InternVL3-78B	0.538	87.6%	0.445	0.058	1.012	1.052	0.768
Sidecar (1 call + OCR)							
Gemini-3.5-Flash + OCR	0.452	91.8%	0.546	0.026	0.928	0.958	0.712
GPT-5.6-Sol + OCR	**0.436**	**92.6%**	**0.578**	**0.018**	**0.912**	**0.918**	**0.698**
Qwen-VL-Max + OCR	0.516	89.1%	0.470	0.050	0.972	0.962	0.778
GLM-4.6V + OCR	0.618	83.8%	0.472	0.088	1.048	1.088	0.792
Doubao-Seed-2.0-pro + OCR	0.596	84.6%	0.584	-0.098	0.862	0.898	0.738
Nova-Pro + OCR	0.459	91.2%	0.538	0.022	0.928	0.902	0.708
Llama-4-Maverick + OCR	0.474	90.4%	0.524	0.030	0.948	0.938	0.718
Pixtral-Large + OCR	0.490	89.6%	0.508	0.036	0.962	0.958	0.732
InternVL3-78B + OCR	0.524	88.2%	0.458	0.052	0.998	1.038	0.758
SFQ-Agent (3 calls)							
Gemini-3.5-Flash	0.440	92.6%	0.560	0.020	0.914	0.942	0.698
Gemini-3.1-Pro	0.428	92.9%	0.572	0.016	0.906	0.932	0.692
GPT-5.6-Sol	**0.418**	**93.4%**	**0.598**	**0.012**	**0.892**	**0.896**	**0.678**
Claude-Sonnet-5	0.456	91.2%	0.542	0.028	0.948	0.842	0.718
Qwen-VL-Max+Qwen-Plus	0.511	89.3%	0.519	0.030	1.007	0.833	0.938
GLM-4.6V	0.574	86.4%	0.508	0.072	1.012	1.052	0.768
Llama-4-Maverick	0.462	91.0%	0.532	0.022	0.942	0.932	0.712
Claude-Opus-4.8	0.424	93.1%	0.588	0.014	0.898	0.882	0.686
Nova-Pro	0.446	92.2%	0.548	0.018	0.924	0.906	0.702

GLM-4.6V lags with the largest errors across dimensions. Sidecar Ablation. On every backend that has a Sidecar row, adding PaddleOCR-VL and CV features reduces MAE, with the largest drop on GLM (ΔMAE$=0.044$) and modest CC/CTX improvements for GPT-5.6-Sol (CC $0.922\rightarrow0.912$, CTX $0.928\rightarrow0.918$). Sidecar is not run for Claude backends, so the claim holds only over the nine Sidecar rows. SFQ-Agent Ablation. On matched backends, staged judging lowers overall MAE relative to Direct, at three calls per figure rather than one. GPT-5.6-Sol has the lowest observed overall error in the frozen matrix (MAE 0.418, W-1 93.4%, SRCC 0.598, bias $+0.012$) and the lowest observed MR error (0.678), above the constant-mean floor (MAE 0.565, W-1 87.9%), for a Direct-best to Agent-best point-estimate gap of 0.025 MAE. We do not attach a significance claim to this gap. Among matched Agent runs, Claude-Sonnet-5 attains the lowest CTX MAE (0.842). The heterogeneous F5 run has a still lower CTX MAE (0.833) but raises MR MAE to 0.938 (Direct Qwen 0.758), so that pairing does not by itself resolve visual and text conflicts.

4.3 Human Ceiling

Table 2 compares models to *mean* gold, whereas Table 3 reports the rater ceiling on the same split and does not restate model MAE: human MAE is leave-one-out versus the other raters, W-1 is pairwise, and α is ordinal Krippendorff. Overall human pairwise W-1 is 71.5% (α=0.51), so Agent W-1 93.4% in Table 2 is versus mean gold, not versus a rater. CC is the unstable construct (pairwise W-1 52.5%, α=0.49, LOO MAE 1.50), and MAE/W-1 versus the mean therefore cannot be read as superhuman caption verification.

Table 3: Human agreement ceiling on *eval1200* (no model scores). MAE: leave-one-out. W-1: pairwise. α: ordinal Krippendorff. CTX: n=804 with $\geq$2 ratings.

Dim.	MAE↓	W-1↑	α↑
VC	0.87	76.6%	0.44
SL	0.62	84.8%	0.42
CC	1.50	52.5%	0.49
CTX	0.88	77.7%	0.61
MR	1.12	67.9%	0.45
Overall	0.70	71.5%	0.51

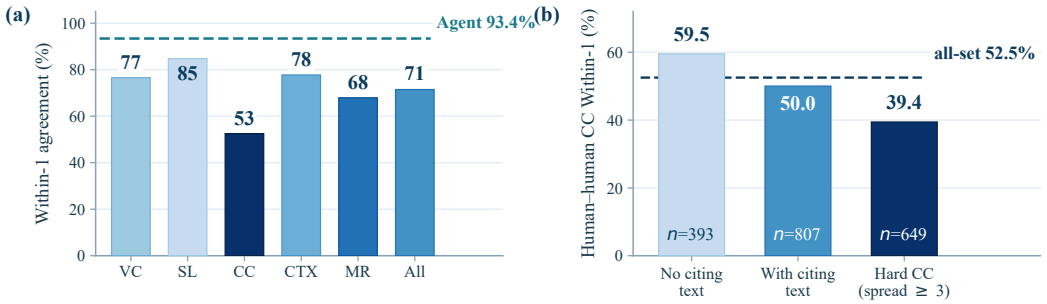


Figure 5: (a) Pairwise human Within-1 by dimension (Table 3), not the same quantity as model W-1. Dashed line: best Agent overall W-1 versus *mean* gold in Table 2 (93.4%). (b) Pairwise human CC Within-1 is lower with citing text (50.0%, n=807) than without (59.5%, n=393), and 39.4% on items with CC spread ≥ 3 (n=649). The with/without split is observational.

Figure 5(a) plots that pairwise ceiling, with the dashed line showing Agent W-1 versus mean gold, a different quantity. Figure 5(b) tests citing text on gold labels: human CC Within-1 is lower with citing text (50.0%, n=807) than without (59.5%, n=393). The split is observational: figures with citing text may already be harder. $\mathcal{T}$ is not a shortcut for caption agreement, even though index-resolved $\mathcal{T}$ remains necessary to score CTX.

4.4 Analysis and Discussion

Where a backend is run under all three protocols, overall MAE falls from Direct to Sidecar to Agent, but Agent uses three calls per figure and the Direct-best to Agent-best gap is only 0.025. Sidecar supplies OCR and layout cues in one call. Agent separates vision from language and lets the Runner lock VC/SL and cap CC, CTX, and MR. We do not claim the gap is independent of model class or compute. The drop is uneven across backends. GLM-4.6V falls from Direct MAE 0.662 to Agent MAE 0.574, while Qwen-VL-Max still has CC MAE 1.007 under Agent. A constant predictor of the *eval1200* mean 8.13 already reaches MAE 0.565 and Within-1 87.9%. Gold scores sit near 8, so Within-1 barely separates protocols. Agent Within-1 93.4% is fit to mean gold. Pairwise human overall Within-1 is 71.5%, and CC remains the least reliable construct (pairwise Within-1 52.5%, α=0.49). These two Within-1 rates ask different questions. 93.4% is how often a model score falls within one point of the rater mean, and 71.5% is how often two raters fall within one point of each other. Citing text does not close the caption gap. Human CC Within-1 is 50.0% when a citing paragraph is present (n=807) and 59.5% when it is absent (n=393). Cited figures may already be harder, so the split is observational. A citing paragraph makes CTX evaluable. It does not supply the panel names, axis titles, or legend entries that CC checks in the caption. Appendix F reads the same split inside five articles, each with every released figure. Overall gold stays near 8.13, while CC moves when the caption names the panels, axes, or legend entries and when it does not. CTX can stay high on the same figure if the citing paragraph states the claim the caption leaves out. The open problem is caption and figure verification under manuscript-linked context, not a small MAE protocol delta. Constant-mean floors and backend-specific behavior are in Appendix E.

5 Conclusion

We introduce SciFigQual-Bench, a manuscript-linked-context benchmark for published CS figures, and SFQ-Agent, a staged protocol that fuses visual, caption, and citing-text evidence into auditable five-dimensional scores. On *eval1200*, SFQ-Agent (F3) has the lowest observed MAE point estimate among Direct, Sidecar, and the constant-mean baseline versus mean gold. By anchoring each figure to its caption and citing paragraphs, the framework offers a reproducible testbed that shifts AI-assisted figure inspection from isolated visual QA toward context-aware scientific reasoning. Caption and figure verification remains the main gap, and model Within-1 should not replace human raters.

AI USE STATEMENT

The experimental protocols query third-party vision and language models as judges (Direct, Sidecar, and SFQ-Agent backends). That usage is the object of study and is reported in Sections 3.4 and 4, not an undeclared authoring aid.

We used generative AI tools to edit wording for readability. All technical claims, numerical results, figures, and bibliographic records were checked by the authors, and we take responsibility for the final content of this work.

ETHICS STATEMENT

This work studies figures that already appear in publicly available CS conference PDFs (ACL, EMNLP, ICML, NeurIPS, 2020 to 2025) and does not collect private user data or scrape non-public manuscripts. Trained human annotators scored figure quality under a written rubric, inspecting published figures, captions, and citing excerpts rather than personal or sensitive attributes. Their professional credentials were not collected or disclosed, so we do not describe the ratings as expert judgments. No identifiable human-subject images or private demographic data are released. Public availability does not by itself grant redistribution rights. The release therefore provides paper identifiers, hashes, figure indices, extraction coordinates, derived annotations, and reconstruction code for all records. Figure crops and text excerpts are redistributed only where the source license permits, otherwise users reconstruct them from the publisher or author-hosted source. A potential misuse is to treat automated figure scores as a sole accept/reject signal in peer review. SFQ-Agent is intended as an assistive, auditable diagnostic, not a replacement for human reviewers. We will release original annotations and code under a research-use license upon acceptance, subject to the source-paper licensing constraints above, and an anonymized supplement is provided for review.

REPRODUCIBILITY STATEMENT

We take the following steps so that the benchmark and judge protocols can be reproduced. Section 3.1 defines the instance tuple, five-dimensional rubric, and L1 gating (Eq. 1). Appendix A.1 and Appendix A.2 give operational anchors and the evaluability mask. Section 3.2 and Appendix A.3 specify corpus acquisition, extraction, curation thresholds, and index-driven context binding. Appendix A.4 records prompt payload limits. Appendix B.1 defines the frozen *eval1200* split. Appendix B.2 maps run IDs. Table 2 is the frozen 29-run matrix used for all model MAE/W-1 in the main text. Appendix B.4 defines the human-ceiling metrics. Table 3 and Figure 5 use gold annotations in `eval1200/figures.jsonl` and do not recompute model scores from that file. Protocol coverage, constant-mean floors, and split composition are in Appendix B.3 and Appendix B.4. Section 3.4 and Appendix C describe the three protocols. Algorithm A1 is the deterministic Runner that locks score ownership and MR fusion. Appendix D documents rater workflow, gold aggregation (Eqs. 3 and 4), and inter-rater dispersion. Appendix F reports five further article-level figure sets with per-rater scores. An anonymized code and data supplement is provided for review.

REFERENCES

Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. TOPIQ: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 33:2404–2418, 2024. doi: 10.1109/TIP.2024.3378466. URL https://doi.org/10.1109/TIP.2024.3378466.

William S. Cleveland and Robert McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984. doi: 10.1080/01621459.1984.10478080. URL https://doi.org/10.1080/01621459.1984.10478080.

Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, Yue Zhang, Yubo Zhang, Handong Zheng, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. PaddleOCR-VL: Boosting multilingual document parsing via a 0.9b ultra-compact vision-language model, 2025. URL https://arxiv.org/abs/2510.14528.

Junpeng Ding, Zichen Tang, Haihong E, Mengyuan Ji, Yang Liu, Haolin Tian, Haiyang Sun, Pengqi Sun, Yang Xu, Yichen Liu, Haocheng Gao, Zijie Xi, Ruomeng Jiang, Peizhi Zhao, Rongjin Li, Yuanze Li, Jiacheng Liu, Zhongjun Yang, Jintong Chen, and Siying Lin. Decoding scientific experimental images: The SPUR benchmark for perception, understanding, and reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12859–12882. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.585. URL https://aclanthology.org/2026.acl-long.585/.

Yaohan Guan, Pristina Wang, Najim Dehak, Alan Yuille, Jieneng Chen, and Daniel Khashabi. GENFIG1: Visual summaries of scholarly work as a challenge for vision-language models, 2026. URL https://arxiv.org/abs/2604.04172.

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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20349–20360, 2023. doi: 10.1109/ICCV51070.2023.01866. URL https://openaccess.thecvf.com/content/ICCV2023/html/Hu_TIFA_Accurate_and_Interpretable_Text-to-Image_Faithfulness_Evaluation_with_Question_Answering_ICCV_2023_paper.html.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386/.

K. V. Jobin, Ajoy Mondal, and C. V. Jawahar. DocFigure: A dataset for scientific document figure classification. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 1, pp. 74–79. Institute of Electrical and Electronics Engineers, 2019. doi: 10.1109/ICDARW.2019.00018. URL https://doi.org/10.1109/ICDARW.2019.00018.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. DVQA: Understanding data visualizations via question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5648–5656, 2018. doi: 10.1109/CVPR.2018.00592. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Kafle_DVQA_Understanding_Data_CVPR_2018_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*, volume 36, pp. 36652–36663, 2023. doi: 10.52202/075280-1594. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/73aacd8b3b05b4b503d58310b523553c-Abstract-Conference.html.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. VIEScore: Towards explainable metrics for conditional image synthesis evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12268–12290. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.663. URL https://aclanthology.org/2024.acl-long.663/.

Seongyun Lee, Seungone Kim, Sue Hyun Park, Geewook Kim, and Minjoon Seo. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11286–11315. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.672. URL https://aclanthology.org/2024.findings-acl.672/.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal ArXiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14369–14387. Association for Computational Linguistics, 2024.

doi: 10.18653/v1/2024.acl-long.775. URL https://aclanthology.org/2024.acl-long.775/.

Wenzhe Li, Liang Chen, Junying Wang, Yijing Guo, Ye Shen, Farong Wen, Chunyi Li, Zicheng Zhang, and Guangtao Zhai. SIQA: Toward reliable scientific image quality assessment, 2026. URL https://arxiv.org/abs/2603.06700.

Xudong Li, Wenjie Nie, Yan Zhang, Runze Hu, Ke Li, Xiawu Zheng, and Liujuan Cao. Distilling spatially-heterogeneous distortion perception for blind image quality assessment. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2344–2354, 2025. doi: 10.1109/CVPR52734.2025.00224. URL https://openaccess.thecvf.com/content/CVPR2025/html/Li_Distilling_Spatially-Heterogeneous_Distortion_Perception_for_Blind_Image_Quality_Assessment_CVPR_2025_paper.html.

Fangyu Liu, Julian Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. Deplot: One-shot visual language reasoning by plot-to-table translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 10381–10399. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.660. URL https://aclanthology.org/2023.findings-acl.660/.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177/.

Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. UniChart: A universal vision-language pretrained model for chart comprehension and reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14662–14684. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.906. URL https://aclanthology.org/2023.emnlp-main.906/.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A dataset for VQA on document images. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 2199–2208, 2021. doi: 10.1109/WACV48630.2021.00225. URL https://openaccess.thecvf.com/content/WACV2021/html/Mathew_DocVQA_A_Dataset_for_VQA_on_Document_Images_WACV_2021_paper.html.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. PlotQA: Reasoning over scientific plots. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1516–1525, 2020. doi: 10.1109/WACV45572.2020.9093523. URL https://openaccess.thecvf.com/content_WACV_2020/html/Methani_PlotQA_Reasoning_over_Scientific_Plots_WACV_2020_paper.html.

Anish Mittal, Anush Krishna Moorthy, and Alan C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. doi: 10.1109/TIP.2012.2214050. URL https://doi.org/10.1109/TIP.2012.2214050.

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. doi: 10.1109/LSP.2012.2227726. URL https://doi.org/10.1109/LSP.2012.2227726.

Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. PieAPP: Perceptual image-error assessment through pairwise preference. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1808–1817, 2018. doi: 10.1109/CVPR.2018.00194. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Prashnani_PieAPP_Perceptual_Image-Error_CVPR_2018_paper.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya

Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Jonathan Roberts, Kai Han, Neil Houlsby, and Samuel Albanie. SciFIBench: Benchmarking large multimodal models for scientific figure interpretation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 18695–18728, 2024. doi: 10.52202/079017-0593. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/217bb44ab14621754db8a392163e6b07-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

Johannes Rückert, Louise Bloch, and Christoph M. Friedrich. Evaluating compliance with visualization guidelines in diagrams for scientific publications using large vision language models. In *Document Analysis and Recognition – ICDAR 2025: 19th International Conference, Wuhan, China, September 16–21, 2025, Proceedings, Part V*, volume 16027 of *Lecture Notes in Computer Science*, pp. 3–21. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-032-04630-7_1. URL https://doi.org/10.1007/978-3-032-04630-7_1.

Noah Siegel, Zachary Horvitz, Roie Levin, Santosh Divvala, and Ali Farhadi. FigureSeer: Parsing result-figures in research papers. In *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII*, volume 9911 of *Lecture Notes in Computer Science*, pp. 664–680. Springer International Publishing, 2016. doi: 10.1007/978-3-319-46478-7_41. URL https://doi.org/10.1007/978-3-319-46478-7_41.

Jifeng Song, Arun Das, Ge Cui, and Yufei Huang. FigEx: Aligned extraction of scientific figures and captions. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 16558–16571. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.899. URL https://aclanthology.org/2025.findings-emnlp.899/.

Edward R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, Connecticut, 2 edition, 2001. ISBN 978-0-9613921-4-7.

He Wang, Longteng Guo, Pengkang Huo, Xuanxu Lin, Yichen Yuan, Jie Jiang, and Jing Liu. S1-MMAlign: A large-scale, multi-disciplinary dataset for scientific figure-text understanding, 2026. URL https://arxiv.org/abs/2601.00264.

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):2555–2563, 2023. doi: 10.1609/aaai.v37i2.25353. URL https://ojs.aaai.org/index.php/AAAI/article/view/25353.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. CharXiv: Charting gaps in realistic chart understanding in multimodal LLMs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 113569–113697, 2024. doi: 10.52202/079017-3609. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/cdf6f8e9fd9aeaf79b6024caec24f15b-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

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 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54015–54029, 2024. URL https://proceedings.mlr.press/v235/wu24ah.html.

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, 2023. URL https://arxiv.org/abs/2306.09341.

Yupeng Xie, Zhiyang Zhang, Yifan Wu, Sirong Lu, Jiayi Zhang, Zhaoyang Yu, Jinlin Wang, Sirui Hong, Bang Liu, Chenglin Wu, and Yuyu Luo. VisJudge-Bench: Aesthetics and quality assessment of visualizations. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=lG1HWWdEbN.

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*, volume 36, pp. 15903–15935, 2023. doi: 10.52202/075280-0700. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/33646ef0ed554145eab65f6250fab0c9-Abstract-Conference.html.

Zhishen Yang, Raj Dabre, Hideki Tanaka, and Naoaki Okazaki. SciCap+: A knowledge augmented dataset to study the challenges of scientific figure captioning. *Journal of Natural Language Processing*, 31(3):1140–1165, 2024. doi: 10.5715/jnlp.31.1140. URL https://doi.org/10.5715/jnlp.31.1140.

Guanghui Ye, Huan Zhao, Zhixue Zhao, Tengfei Ma, Kehan Wang, Steffen Eger, and Zhihua Jiang. SCIEval: Evaluating and benchmarking the faithfulness of scientific image generation and interpretation with large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29760–29770, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Ye_SCIEval_Evaluating_and_Benchmarking_the_Faithfulness_of_Scientific_Image_Generation_CVPR_2026_paper.html.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html. Datasets and Benchmarks Track.

Supplementary Material

Appendix

A Benchmark Construction Details

A.1 Five-Dimensional Rubric

In SciFigQual-Bench, all manual scoring rules are uniform: every score is an integer on $[1, 10]$, VC, SL, CC, and CTX increase with quality, and MR uses the same range with positive polarity (higher means a careful reader is less likely to be misled). Dimensions are independent, so a sharp figure can still fail CC/CTX and a weak caption must not automatically lower VC/SL. Direct, Sidecar, and SFQ-Agent reuse one rubric text and differ only in how evidence is collected. Humans and all models share that rubric, with only the evidence path changing.

All human raters and models follow the same bands: 9 to 10 means near-absence of relevant defects, 7 to 8 minor defects that do not block the main reading, 5 to 6 noticeable friction, 3 to 4 serious impedance, and 1 to 2 near-unusable or severely misleading content. On CC, CTX, and MR, unresolved ties between adjacent scores are resolved toward the lower score. The 5 to 8 bands describe mixed-quality cases in the rubric, while empirical scores on this published-only pool concentrate near 8 (Table A12).

Table A1 lists operational anchors. VC checks publication-scale legibility (blur, contrast, ticks, legends). SL checks reading path, panel labels, spacing, and chartjunk (Tufte, 2001). CC checks whether caption c names the right objects, metrics, panels, and trends. CTX checks whether citing paragraphs $\mathcal{T}$ make claims that I supports, and is not a writing-quality score. MR checks truncated axes, missing baselines/error bars, unfair comparison, and visual and text conflicts (Rückert et al., 2025), kept separate from CC/CTX so that a vague caption alone does not collapse trust when the plot is fair.

Table A1: Operational anchors for the five scoring dimensions.

Dim.	Primary cues	High band	Low band
VC	Resolution, blur, axis/legend legibility	Readable at publication scale	Effortful (5 to 6). Illegible key labels (3 to 4). Unreadable (1 to 2)
SL	Whitespace, panel order, $(a)/(b)$, flow	Clear path and complete labels	Costly structure (5 to 6). Disordered (3 to 4). No hierarchy (1 to 2)
CC	Entities, metrics, panels, numeric alignment	Covers objects/metrics/panels	Vague (5 to 6). Mismatch (3 to 4). Wrong-figure (1 to 2)
CTX	Index, trends/comparisons vs. I	Specific evidentiary match	Bare pointer (5 to 6). Wrong claim/figure (3 to 4 / 1 to 2)
MR	Axes, baselines, fairness, conflicts	Trustworthy and fair	Ambiguous (5 to 6). Missing signals (3 to 4). Contradiction (1 to 2)

To prevent fluent but content-empty text from receiving inflated scores, we impose hard caps on CC and CTX. A caption that describes the figure but omits the key metric or compared groups receives CC at most 5 to 6 (cap $= 6$), and a citation that merely points to the figure without articulating any finding (e.g., "see Figure 3") is capped at 5 to 6 on CTX, never above 6. By contrast, a citation that reverses the actual trend receives 1 to 2 on both CTX and MR. These caps apply independently of visual quality: a visually pristine figure cannot compensate for substantive textual misalignment, OCR- or CV-derived evidence may support VC/SL but cannot override the pixel-level content, and inapplicable sub-criteria are ignored.

A.2 L1 Evidence Gating

Missing metadata must not be scored as low quality. Table A2 is the mask shared by human gold and every model protocol: if both c and $\mathcal{T}$ are empty the instance is dropped. Otherwise VC/SL/MR stay available from the image, CC requires a caption, CTX requires citing text, and the overall is the arithmetic mean of the remaining dimension scores.

Table A2: L1 gating of evaluable dimensions.

Evidence	VC	SL	CC	CTX	MR
c and $\mathcal{T}$ present	yes	yes	yes	yes	yes
only c	yes	yes	yes	null	yes
only $\mathcal{T}$	yes	yes	null	yes	yes
both absent	instance excluded				

A.3 Curation Stages S1 and S5

Raw layout extraction yields many non-evaluable crops. We therefore run five sequential stages that separate hard deletion from soft retention (Table A3). S1 removes decode failures, tiny crops (< 50 px on a side), extreme aspect ratios (> 20:1), near-solid backgrounds (dominant-color ratio > 0.97), and near-duplicate pHashes (Hamming ≤ 8). S2 only *flags* anomalies within a paper (area $> 3\times$ the paper median, height or width overflow, or an abnormal figure index), without deleting rows. S3 re-crops from the PDF with caption-anchored boxes when flags indicate broken extraction. Failed recoveries fall back to the original crop and are marked rather than silently dropped. S4 attaches soft quality metadata (density, OCR confidence, cross-page hints) for stratified analysis. S5 drops papers whose unusable-figure rates exceed fixed thresholds (e.g., hard-delete ratio > 0.30), so context binding is not attempted on irrecoverable layouts. A figure must survive S1 and S3 to enter the released set. S4 keeps borderline cases. S5 acts at paper granularity. The funnel from K_0=62,694 PDFs to P=1,144 papers and N=7,609 figures is the result of these filters, not random subsampling.

Table A3: Curation stages and primary actions (implementation thresholds).

S	Name	Primary action
1	Hard filter	Delete decode failures, < 50 px, aspect > 20, solid > 0.97, pHash dup. ≤ 8
2	Anomaly flag	Soft-flag oversized / overflow / bad-index crops (no delete)
3	Re-extract	Caption-anchored PDF re-crop for flagged failures. Else keep original
4	Soft mark	Attach quality metadata for later stratification
5	Paper filter	Drop papers with repeated unusable figures / missing indices

A.4 Context Binding and Prompt Payloads

Citing bundles $\mathcal{T}_k$ are resolved from figure-index patterns in body text (e.g., "Figure N" / "Fig. N"), not from abstract heuristics. Matched paragraphs are merged within section boundaries. Blocks shorter than 20 characters are ignored. For scoring prompts we further truncate payloads so that backends see comparable context: up to two snippets of at most 500 characters for the target figure, plus an optional paper-level figure map with shortened captions (≤ 220 characters). After aggregation, bare-pointer paragraphs are down-weighted and comparative/trend language is preferred for CTX, but we never invent citations. Absent index matches leave CTX gated null. Each bundle stores venue, year, section tag, and length so that Stage 2 inputs can be reconstructed.

B Experimental Protocol Details

B.1 eval1200 Split Definition

eval1200 is the frozen evaluation set for all protocol and backend ablations (n=1,200 figures from the 6,308 rated instances), and each figure has at least three independent human ratings (Table A11). Gold labels and protocol contracts were frozen before any model call. The split is released publicly for reproducibility after that freeze. It is not presented as a permanently hidden leaderboard test. Sampling is paper-aware (figures from the same paper are kept or dropped together), yielding 1,200 figures from 254 papers mixed by figure type rather than forced to equal venue counts. Table A4 records venue and year counts and Table A5 records topic labels on the same split. Gold labels are frozen before any model call. Every Direct / Sidecar / SFQ-Agent configuration consumes identical $(I, c, \mathcal{T}, m)$ under the L1 mask in Table A2. The 29 rows are not a complete 11×3 factorial: Sidecar omits Claude backends. Agent adds Gemini-3.1-Pro and omits Doubao, Pixtral, and InternVL (Table A6).

Table A4: *eval1200* venue and year composition (n=1,200 figures, 254 papers). Venue counts match the main-text split. Years are from the gold JSONL.

Venue	Figures
ACL	266
EMNLP	294
ICML	294
NeurIPS	346
All	1,200

Year	Figures
2020	155
2021	140
2022	197
2023	190
2024	272
2025	246
All	1,200

Table A5: Topic labels on *eval1200* (n=1,200). "Other labeled" aggregates reinforcement learning, multimodal, graph learning, and speech. Unlabeled items have no topic tag in the release.

Topic	Figures	Share
Natural language processing	535	44.6%
Machine learning (general)	360	30.0%
Computer vision	139	11.6%
Other labeled CS topics	77	6.4%
Unlabeled	89	7.4%
All	1,200	100%

B.2 Run ID Mapping

Each configuration has a stable run ID and call budget. Direct (1 call/figure): D1 Gemini-3.5-Flash, D2 GPT-5.6-Sol, D3 Claude-Sonnet-5, D4 Qwen-VL-Max, D5 GLM-4.6V, D6 Doubao-Seed-2.0-pro, D7 Llama-4-Maverick, D8 Pixtral-Large, D9 Nova-Pro, D10 Claude-Opus-4.8, D11 InternVL3-78B. Sidecar (1 call/figure): S1 Gemini-3.5-Flash, S2 GPT-5.6-Sol, S3 Qwen-VL-Max, S4 GLM-4.6V, S5 Doubao-Seed-2.0-pro, S6 Llama-4-Maverick, S7 Pixtral-Large, S8 Nova-Pro, S9 InternVL3-78B, each with PaddleOCR-VL (Cui et al., 2025) side features. SFQ-Agent (3 calls/figure): F1 Gemini-3.5-Flash/Gemini-3.5-Flash, F2 Gemini-3.1-Pro/Gemini-3.1-Pro, F3 GPT-5.6-Sol/GPT-5.6-Sol, F4 Claude-Sonnet-5/Claude-Sonnet-5, F5 Qwen-VL-Max+Qwen-Plus, F6 GLM-4.6V/GLM-4.6V, F7 Llama-4-Maverick/Llama-4-Maverick, F8 Claude-Opus-4.8/Claude-Opus-4.8, F9 Nova-Pro/Nova-Pro. Slash notation denotes matched vision/language backends. F5 is the only heterogeneous pairing. Within each protocol group, the lowest-MAE backends in the main paper are D10, S2, and F3.

B.3 Protocol and Backend Coverage

The 29-run MAE/W-1 matrix is Table 2 in the main text (the frozen copy used for all model scores there). Table A6 records which backends were run under each protocol. Empty cells are not imputed. Direct and Sidecar use one call per figure. Agent uses three.

Table A6: Run-ID coverage of the 29 configurations. Empty cells were not run. F5 is the only heterogeneous vision/language pairing (Qwen-VL-Max + Qwen-Plus).

Backend	Direct	Sidecar	Agent
Gemini-3.5-Flash	D1	S1	F1
Gemini-3.1-Pro	n/a	n/a	F2
GPT-5.6-Sol	D2	S2	F3
Claude-Sonnet-5	D3	n/a	F4
Qwen-VL-Max	D4	S3	F5
GLM-4.6V	D5	S4	F6
Doubao-Seed-2.0-pro	D6	S5	n/a
Llama-4-Maverick	D7	S6	F7
Pixtral-Large	D8	S7	n/a
Nova-Pro	D9	S8	F9
Claude-Opus-4.8	D10	n/a	F8
InternVL3-78B	D11	S9	n/a

B.4 Human-Ceiling and Slice Metrics

Table 3 and Figure 5 use gold annotations in `eval1200/figures.jsonl` only. They do not restate model MAE/W-1 from Table 2. Table A7 distinguishes pairwise human agreement from model agreement versus mean gold. Human leave-one-out MAE compares each rater to the mean of the other raters on that figure and dimension. Pairwise Within-1 averages all rater pairs. Ordinal Krippendorff's α uses coincidence counts on integer scores in $\{1, \ldots, 10\}$. Hard CC is the subset whose per-figure CC spread is at least 3 points (n=649). Of 807 figures with citing text, 804 retain $\geq$2 CTX ratings after per-rater nulls (three keep only one; Table 3). Figure 5(b) reports the 807/393 citing split, restated numerically in Table A10. The dashed line in Figure 5(a) is the best Agent overall W-1 from Table 2 (93.4%), which is versus *mean* gold rather than pairwise human agreement. A constant predictor of each dimension's *eval1200* mean is reported in Table A8. Overall MAE is 0.565 and W-1 is 87.9% versus mean 8.13. Table A9 places that floor next to the best row of each protocol in Table 2.

Table A7: Metric definitions used in the main text. Model scores always follow Table 2. Human-ceiling numbers follow gold JSONL only.

Quantity	Definition	Where used
Pairwise W-1	Fraction of rater pairs with $\lvert s_r - s_{r'} \rvert \leq 1$	Table 3. Figure 5
Leave-one-out MAE	$\lvert s_r - \text{mean}_{r' \neq r} s_{r'} \rvert$ averaged over raters and items	Table 3
Ordinal α	Krippendorff α on integer scores $\{1, \ldots, 10\}$. Items with < 2 ratings dropped	Table 3
Model MAE / W-1 / SRCC	Versus dimension-first *mean* gold (Eqs. 3 and 4)	Table 2
Constant MAE / W-1	Predict the *eval1200* mean of that dimension	Table A8

C Judge Prompts and Deterministic Runner

Direct and Sidecar share a single-pass template. Sidecar additionally receives OCR/CV features. SFQ-Agent issues three staged calls, then applies the code-level Runner in Algorithm A1. Full prompt templates are included in the anonymized code supplement.

Table A8: Constant-mean predictor on *eval1200* gold (no model outputs). Each row predicts that dimension's empirical mean. CTX is restricted to the 807 figures with citing text. SRCC is undefined for a constant.

Dim.	Mean	n	MAE↓	W-1↑
VC	8.31	1,200	0.572	83.3%
SL	8.61	1,200	0.404	95.4%
CC	7.48	1,200	1.092	53.2%
CTX	8.27	807	0.989	60.2%
MR	8.04	1,200	0.778	79.4%
Overall	8.13	1,200	0.565	87.9%

Table A9: Overall *eval1200* constant floor versus the best row of each protocol in Table 2. Constant and model W-1 are both versus *mean* gold, not pairwise human W-1.

Predictor	MAE↓	W-1↑	SRCC↑
Constant mean (8.13)	0.565	87.9%	n/a
Direct-best (D10)	0.443	92.4%	0.582
Sidecar-best (S2)	0.436	92.6%	0.578
Agent-best (F3)	0.418	93.4%	0.598

C.1 Single-Pass Judges (Direct and Sidecar)

Direct: one VLM call over $(I, c, \mathcal{T})$ returns five integer scores with short evidence-tied reasons. Gated-off CC/CTX are `null`. Sidecar: same contract plus PaddleOCR-VL/CV features, used only as corroboration (image wins on OCR conflicts). Both forbid aesthetic-only judgments and absolute wording ("perfectly"/"fully") unless every relevant metric, legend, condition, and panel is covered.

C.2 SFQ-Agent Evidence Stages

Vision. VLM reads *I*+features, *writes final* VC/SL, and emits image facts plus visual risk band $[\ell, h]$. It must not score CC/CTX/final MR.

Language. LLM reads only *c*/$\mathcal{T}$ (no pixels) and emits *text-side* proposals: omissions/bare-reference caps $\mathrm{cap}_{\mathrm{CC}}, \mathrm{cap}_{\mathrm{CTX}}$ (typically 6) and severity $\sigma \in \{\texttt{none}, \texttt{minor}, \texttt{moderate}, \texttt{severe}, \texttt{contradiction}\}$ as a prior over caption/context friction, not a pixel-verified cross-modal verdict. No final scores.

Fusion. Judge proposes CC/CTX and an MR *rationale* from the two evidence reports (any numeric MR it emits is only a draft). The Runner then *overwrites* VC/SL from Stage 1 and *replaces* the MR number by Algorithm A1's closed form of $[\ell, h]$ and σ, while applying $\mathrm{cap}_{\mathrm{CC/CTX}}$ and optional $\mathrm{cap}_{\mathrm{MR}}$ (Stage 2 `misleading_score_cap`, usually `null`, and set only for an explicit severe contradiction that must not receive a high MR).

C.3 Runner Hard Constraints

The non-prompt part of SFQ-Agent is Algorithm A1: hard-copy VC/SL from the VLM, clamp CC/CTX (and CTX under σ=`contradiction`), and replace MR by a closed-form function of $[\ell, h]$ and σ. Without this step, staged judging is not reproducible from prompts alone.

Table A10: Pairwise human CC Within-1 on *eval1200* slices (gold only). Companion to Figure 5(b). The with/without citing-text split is observational.

Slice	n	Pairwise CC W-1
All *eval1200*	1,200	52.5%
No citing text	393	59.5%
With citing text	807	50.0%
Hard CC (spread ≥ 3)	649	39.4%

Algorithm A1 SFQ-Agent Runner: ownership, caps, and MR fusion.

Require: scores s. Band $[\ell, h]$. Severity σ. Caps $\mathrm{cap}_{\mathrm{CC}}$, $\mathrm{cap}_{\mathrm{CTX}}$ (opt. $\mathrm{cap}_{\mathrm{MR}}$ from Stage 2)

$s(\mathrm{VC}) \leftarrow s_{\mathrm{VLM}}(\mathrm{VC})$. $s(\mathrm{SL}) \leftarrow s_{\mathrm{VLM}}(\mathrm{SL})$

if CC on **and** $\mathrm{cap}_{\mathrm{CC}}$ set **then** $s(\mathrm{CC}) \leftarrow \min(s(\mathrm{CC}), \mathrm{cap}_{\mathrm{CC}})$

if CTX on **and** $\mathrm{cap}_{\mathrm{CTX}}$ set **then** $s(\mathrm{CTX}) \leftarrow \min(s(\mathrm{CTX}), \mathrm{cap}_{\mathrm{CTX}})$

if CTX on **and** σ=`contradiction` **then** $s(\mathrm{CTX}) \leftarrow \min(s(\mathrm{CTX}), 2)$

$m \leftarrow (\ell + h)/2$

$$r \leftarrow \begin{cases} \texttt{none}: & m \\ \texttt{minor}: & \max(\ell, m-1) \\ \texttt{moderate}: & \min(m, 6.5) \\ \texttt{severe}: & \min(m, 4.5) \\ \texttt{contradiction}: & 2 \end{cases}$$

$s(\mathrm{MR}) \leftarrow \mathrm{clip}_{[1,10]}(\mathrm{round}(r))$

if $\mathrm{cap}_{\mathrm{MR}}$ set **then** $s(\mathrm{MR}) \leftarrow \min(s(\mathrm{MR}), \mathrm{cap}_{\mathrm{MR}})$

return s and L1-gated mean over scored dims

D Human Annotation Protocol

D.1 Rater Workflow

Annotators inspect I first, then c, then $\mathcal{T}$, and only afterwards optional OCR/CV features. Model predictions are never shown, so humans are not anchored by automated scores. Each dimension requires a short rationale grounded in visible marks or exact phrases, and plausible but unfaithful explanations are rejected (Jacovi & Goldberg, 2020). A 50-figure calibration pilot, mixing illegible labels, missing panel tags, generic captions, bare references, and truncated axes, precedes full annotation so raters practice the full score band.

D.2 Quality Control and Dimension Statistics

Dual annotation on an ACL 2025 holdout (n=281 figures, 48 papers) checked rubric operability before scaling, and overall disagreements exceeding two points on that holdout triggered senior adjudication. *eval1200* is distinct from that dual holdout: every figure has at least three ratings, 119 figures have additional ratings (Table A11), and released gold uses all available ratings under Eqs. 3 and 4. Table A12 compares dimension means on the full rated pool and on *eval1200*. CC is the weakest axis on both, because many published figures are visually adequate yet captions omit metrics or comparison groups.

D.3 Gold Standard Aggregation

Released gold is aggregated *dimension-first*, matching the L1 mask at inference. For each evaluable dimension d and figure i,

$$\bar{s}_i(d) = \frac{1}{|\{r : s_{i,r}(d) \neq \mathrm{null}\}|} \sum_r s_{i,r}(d), \tag{3}$$

Table A11: Number of human ratings per *eval1200* figure. Gold means and IRR use every available rating on that figure.

# raters	Figures
3	1,081
4	64
6	1
7	1
8	50
9	3
All	1,200

Table A12: Mean human scores on the full rated pool versus *eval1200*. CTX means are over items with citing text (L1-evaluable).

Dim.	**Rated pool** (n=6,308)	***eval1200*** (n=1,200)
VC	8.12	8.31
SL	8.58	8.61
CC	7.42	7.48
CTX	7.89	8.27
MR	8.01	8.04
Overall	8.05	8.13

then

$$\bar{y}_i = \frac{1}{\sum_d g_i(d)} \sum_d g_i(d)\, \bar{s}_i(d). \tag{4}$$

On a complete rating matrix (every rater scores the same evaluable dimensions), $\bar{y}_i$ equals the mean of per-rater overalls by linearity. The two aggregations diverge only when raters have *uneven* L1 null masks on the same figure (e.g., one rater nulls CTX while another scores it). The release always uses Eqs. 3 and 4.

Worked example (Ex-B in Figure A1). Ex-B receives rater scores $(6, 5, 5, 6, 4)$, $(5, 6, 5, 6, 4)$, and $(8, 9, 6, 7, 5)$. Per-rater overalls are 5.20, 5.20, and 7.00. Dimension means are $\bar{s} = (6.33, 6.67, 5.33, 6.33, 4.33)$, so $\bar{y} = 5.80$, which coincides with the mean of the three per-rater overalls on this complete case. The release uses 5.80.

Null handling. Absent $\mathcal{T}$ nulls CTX for every rater (Ex-A). Absent c nulls CC. Table A13 reports dispersion on *eval1200*: CC has the largest mean spread (2.65) and lowest W-1 pair rate (52.5%), aligning with its role as the hardest axis for both humans and models.

D.4 Representative Multi-Rater Cases

Figure A1 shows three *eval1200* instances chosen to illustrate gating, disagreement, and agreement. Ex-A has a caption but no citing text, so CTX is null and difficulty concentrates on CC. Ex-B lacks clear axis labels. Raters 1 and 2 are substantially harsher on MR than Rater 3 (per-rater overall gap 1.8, below the holdout's two-point adjudication trigger). Ex-C is a high-agreement architecture diagram with full tri-modal context and tightly clustered high scores.

D.5 Human vs. Model Comparison Cases

Figure A2 compares humans with Direct GPT-5.6-Sol (D2), Direct Qwen-VL-Max (D4), and SFQ-Agent GPT-5.6-Sol (F3) on three further instances. Ex-F is already high-agreement. Staging changes little. Ex-G is a caption bottleneck: Direct over-credits fluent but incomplete captions, while F3

Table A13: Inter-rater dispersion on *eval1200* (n=1,200 figures; at least three raters on VC/SL/CC/MR/overall). Spread $= \max_r s_{i,r}(d) - \min_r s_{i,r}(d)$ per figure. Median is the median of that spread. W-1 pair rate = fraction of rater pairs within ± 1 point.

Dimension	Mean spread	Median spread	W-1 pair	n
VC	1.52	1.0	76.6%	1,200
SL	1.03	1.0	84.8%	1,200
CC	2.65	3.0	52.5%	1,200
CTX	1.54	1.0	77.7%	804
MR	1.92	2.0	67.9%	1,200
Overall (per-rater)	1.26	1.0	71.5%	1,200

eval1200 means: VC 8.31. SL 8.61. CC 7.48. CTX 8.27. MR 8.04. Overall 8.13. CTX n=804 items with $\geq$2 ratings (807 figures have citing text).

checks caption facts against image-side expectations and moves toward the human CC/overall region. Ex-H is a training-curve mismatch: single-pass judges soften the contradiction, whereas Runner-backed fusion lowers CC/MR toward the human gold by combining the visual risk band with text severity (Algorithm A1).

D.6 Annotation Interface and Adjudication

Each record stores five dimension scores, a per-rater gated overall, per-dimension reason strings, plus `summary`/`suggestion`. Contested or low scores require reasons. Unfaithful rationales are rejected in review. The full JSONL for all 1,200 *eval1200* instances (with aggregated human means) is included in the anonymized data supplement. On the ACL 2025 holdout, per-rater overall gaps above two points trigger adjudication before release. Released gold then follows Eqs. 3 and 4.

E Extended Analysis and Limitations

Table A14 makes the metric gap concrete: a constant-mean predictor already reaches W-1 87.9%, so Within-1 barely separates protocols, while MAE and SRCC carry more of the signal. Even the highest observed SRCC is 0.598, consistent with gold scores concentrated near 8.

Table A14: Overall *eval1200* floor versus protocol-best rows from Table 2. Model and constant W-1 are both versus *mean* gold, not pairwise human agreement.

Predictor	MAE↓	W-1↑	SRCC↑
Constant mean (8.13)	0.565	87.9%	n/a
Direct-best (Claude-Opus-4.8)	0.443	92.4%	0.582
Sidecar-best (GPT-5.6-Sol)	0.436	92.6%	0.578
Agent-best (GPT-5.6-Sol)	0.418	93.4%	0.598

Backend Behavior. GPT-5.6-Sol is the most balanced, with near-zero bias under Agent. Qwen-VL-Max is calibrated under Direct (W-1 88.8%) but ranks poorly (SRCC 0.416). Staging recovers SRCC to 0.519 while CC MAE stays at 1.007. GLM-4.6V improves most under richer protocols (Direct MAE 0.662 to Agent MAE 0.574), whereas Doubao stays negatively biased (-0.120 to -0.098). These are descriptive comparisons at unequal protocol-level call budgets.

Scope and Limitations. Per-dimension error localizes failure along legibility, caption fit, narrative support, and misleading risk. The protocol ladder is a design choice, not a trained specialist judge, and CC remains a weak construct (pairwise W-1 52.5%, α=0.49 in Table 3): fitting mean gold is easier than matching a rater, and published-only figures compress the score range. CTX is unavailable when no explicit citing paragraph is resolved. The evaluation uses bounded citing snippets rather than full-paper model context, the 29-run matrix is not a complete factorial, Agent uses more calls than Direct, and we do not report paper-clustered confidence intervals or repeated-API variance for the 0.025 MAE gap. Rankings are therefore descriptive point estimates for frozen run IDs,

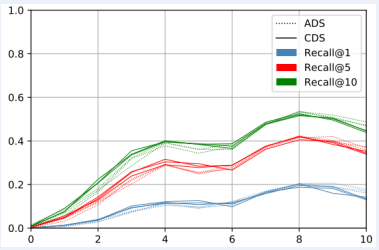
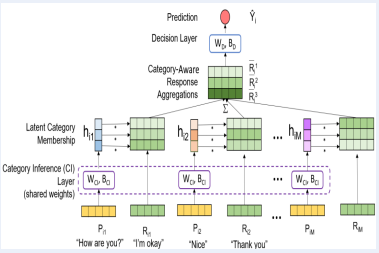
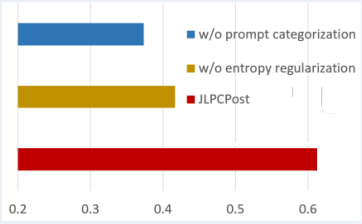
ID	Image path	Caption & citing text	Expert 1	Expert 2	Expert 3	Gold
Ex-A		Cap.: Figure 1: Validation performance in early training on natural speech. Cite: (none; CTX gated null) CC rationale (Expert 1): "caption broadly describes performance but omits Recall@K metrics and ADS/CDS groups visible in the legend."	Expert 1 9/9/5/--/9 = 8.0	Expert 2 8/8/5/--/8 = 7.25	Expert 3 9/9/6/--/7 = 7.75	8.67/8.67/5.33/--/8.00 = 7.67
Ex-B		Cap.: Figure 2: Ablation study on validation set demonstrating the importance of prompt categorization and entropy regularization for our model. Cite: ...the contributions of each component are visualized in Figure 2... latent prompt categorization yields the highest contribution... MR rationale (Expert 2): "absence of a y-axis label and units for the x-axis makes it impossible to fully understand what performance is being measured." MR rationale (Expert 3): "X-axis starting at 0.2 visually exaggerates differences between bars."	Expert 1 6/5/5/6/4 = 5.2	Expert 2 5/6/5/6/4 = 5.2	Expert 3 8/9/6/7/5 = 7.0	6.33/6.67/5.33/6.33/4.33 = 5.80
Ex-C		Cap.: Figure 1: The architecture of our JLPC model with K=3... Category Inference layer... Decision Layer... Cite: Our model, shown in Figure 1, takes as input an interview Xi and outputs the predicted label Y-hat-i... CTX rationale (Expert 3): context explains inputs (Xi), outputs (Y-hat-i), embeddings (Pij, Rij), and latent category vectors (hij), all clearly part of the visual flow.	Expert 1 8/9/9/8/9 = 8.6	Expert 2 9/9/8/9/9 = 8.8	Expert 3 9/9/9/9/8 = 8.8	8.67/9.00/8.67/8.67/8.67 = 8.73

Figure A1: Human multi-rater annotation cases from *eval1200*. Each block is one figure. Scores are on $[1, 10]$ (higher is better; MR means lower misleading risk). Gold column reports dimension-first means. n/a marks gated-null CTX. Rationale snippets are verbatim from annotator records.

not evidence of a universal ordering. Future releases should retain a paper-disjoint hidden test for leaderboard use and report uncertainty clustered by source paper.

F Article-Level Figure Sets

Figure A1 and Figure A2 each show one crop. The five articles below are disjoint from those examples. For each article we keep every released figure in *eval1200*, and the indices run from 1 with no gaps. All five dimensions are scored, so neither CC nor CTX is an L1 null. Gold follows Eqs. 3 and 4. Scores lie on $[1, 10]$. Higher MR means lower misleading risk. Source titles, venues, and years are omitted. Raters are written as R1, R2, and R3 inside each case. A per-rater vector is $(\mathrm{VC}, \mathrm{SL}, \mathrm{CC}, \mathrm{CTX}, \mathrm{MR})$. These rows are human ratings. Model predictions are not attached.

The sets sit next to two results in the main text. On *eval1200*, CC is the weak axis (pairwise W-1 52.5%, α=0.49 in Table 3, mean spread 2.65 and median spread 3.0 in Table A13), while SL is

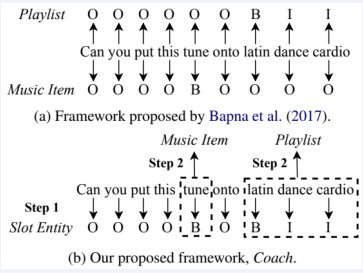
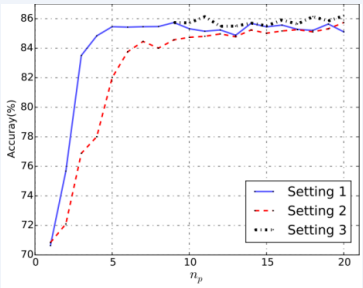
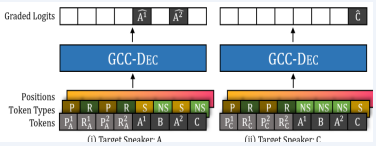
ID	Image path	Caption & context (excerpt)	Human gold	D2 Direct	D4 Direct	F3 Agent
Ex-F		Cap.: Figure 1: Illustration of input representation for a multi-actor conversation... Cite: ...designate a speaker of interest... As visualized in Figure 1, the designated speaker's reference history tokens are gathered... Analysis: Strong tri-modal agreement; F3 tracks gold within +/-0.5 on all dimensions, while D2/D4 slightly over-score overall.	8.33/8.67/9.00/9.00/9.00 = 8.80	9.1/9.0/9.1/8.8/8.9 = 9.1	9.0/8.7/9.2/9.2/9.3 = 9.1	8.3/8.5/8.8/8.8/8.7 = 8.7
Ex-G		Cap.: Figure 1: Cross-domain slot filling frameworks. Cite: As illustrated in Figure 1a, their model conducts slot filling individually for each slot type... Figure 1b, we propose... Coach... Human CC note: caption is overly general—omits Bapna et al. vs. Coach panels (a/b). Model gap: D2/F3 still rate CC 1.1-1.5 points above human gold.	8.67/8.67/5.67/9.00/8.33 = 8.07	8.5/8.4/7.2/8.5/8.9 = 8.5	7.9/7.9/6.9/8.2/8.1 = 7.9	8.1/8.2/6.8/8.0/7.6 = 7.6
Ex-H		Cap.: Fig. 3: Training time and accuracy vs. #FS terms (blue line = accuracy; red bar = training time). Cite: ...training time spent and the final accuracy achieved... in Fig. 3... Human MR note: caption claims a red bar for training time, but only accuracy lines are plotted. Models: D2 over-estimates overall (+1.5); D4 is closest on CC/MR.	7.00/8.00/3.00/5.00/3.00 = 5.20	7.0/7.4/5.0/5.9/4.7 = 6.7	5.5/6.1/3.8/5.0/4.0 = 4.6	6.4/6.7/4.5/5.2/4.2 = 5.6

Figure A2: Human gold vs. model predictions on selected *eval1200* instances (disjoint from Figure A1). Format: VC/SL/CC/CTX/MR = Overall (unweighted mean of the five printed dimensions). Human column is dimension-first gold. Model columns apply the same L1 gating at inference. The Direct prompt requests integers in $[1, 10]$; displayed cells may be one-decimal parses.

tight (mean spread 1.03). Overall gold on *eval1200* is 8.13, and the articles below also land near that band. A short caption can leave CC low even when the citing paragraph carries the claim, which is the caption and figure check under manuscript-linked context, read across the figures of one article.

Case 1. Three figures. Figure 1 is a block diagram. A candidate document enters a news encoder on the left. Clicked documents enter a user encoder on the right. The two vectors meet at a dot product labeled as click probability. The caption names the framework and does not mention the encoders or the predictor. Figure 1 receives $(9, 9, 8, 9, 9)$, $(9, 9, 8, 9, 9)$, and $(8, 9, 5, 7, 7)$, so $\bar{s} = (8.67, 9.00, 7.00, 8.33, 8.33)$ and $\bar{y} = 8.27$. SL is 9 for every rater. CC is $8/8/5$. The score of 5 treats the caption as topically right but too thin for the drawn data flow. The citing paragraph uses the figure to separate this matching setup from a later generation task, so CTX is $9/9/7$ and stays above CC. Visual clarity is not the failure. The caption is.

Figure 2 is six histograms. Panels (a) and (b) are title length and body length. Panels (c) and (d) are entity counts in the title and in the body. Panel (e) is topic share, and panel (f) is click-history length. The topic labels in (e) are small, which is why one rater gives VC 7. The caption says the

Table A15: Case 1. Dimension scores and gated overall.

Fig.	Rater	VC	SL	CC	CTX	MR	Ov.
1	R1	9	9	8	9	9	8.80
1	R2	9	9	8	9	9	8.80
1	R3	8	9	5	7	7	7.20
1	Gold	8.67	9.00	7.00	8.33	8.33	8.27
2	R1	7	8	8	9	9	8.20
2	R2	9	9	6	8	9	8.20
2	R3	9	9	5	9	7	7.80
2	Gold	8.33	8.67	6.33	8.67	8.33	8.07
3	R1	7	8	10	9	9	8.60
3	R2	9	8	9	9	9	8.80
3	R3	9	9	10	10	6	8.80
3	Gold	8.33	8.33	9.67	9.33	8.00	8.73

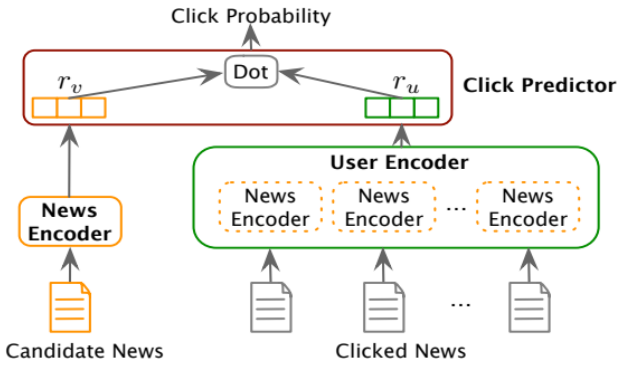
figure shows corpus and training-set statistics, and it does not name the six panels. Figure 2 receives $(7, 8, 8, 9, 9)$, $(9, 9, 6, 8, 9)$, and $(9, 9, 5, 9, 7)$, so $\bar{s} = (8.33, 8.67, 6.33, 8.67, 8.33)$ and $\bar{y} = 8.07$. CC is $8/6/5$. The low CC asks the caption to list the panels. The citing paragraph assigns topics to panel (c) and click history to panel (f). On the drawing, topics are panel (e) and click history is panel (f), so the panel letter for topics does not match. Gold CTX is still 8.67. The same paragraph states a mean title length of 10.5 and a mean body length of 549.0, and both values fall inside the bins of panels (a) and (b). CC gold 6.33 against CTX gold 8.67 is caption underspecification with partial citing-text support, inside one article.

Figure 3 is the generation diagram. Three numbered arrows mark where a user vector enters the decoder. The first initializes the decoder state, the second changes attention over the source words, and the third changes the choice between generating and copying. The caption lists those three routes in the same order as the drawing. Figure 3 receives $(7, 8, 10, 9, 9)$, $(9, 8, 9, 9, 9)$, and $(9, 9, 10, 10, 6)$, so $\bar{s} = (8.33, 8.33, 9.67, 9.33, 8.00)$ and $\bar{y} = 8.73$. CC is $10/9/10$, unlike Figures 1 and 2. VC is 7 for one rater because the numerals 1, 2, and 3 are small and light. One rater scores MR at 6 because blocks such as F_1, F_2, and the copy probability are drawn without their internal computation. The other two score MR at 9. Gold MR is 8.00. Inside this article, overall gold moves from 8.07 to 8.73, near the *eval1200* mean of 8.13, while CC moves from 6.33 to 9.67.

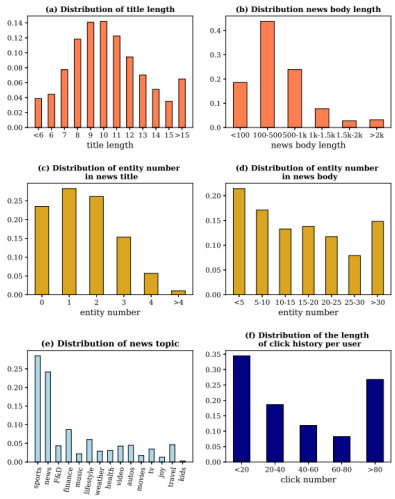
Table A16: Case 2. Dimension scores and gated overall.

Fig.	Rater	VC	SL	CC	CTX	MR	Ov.
1	R1	9	9	10	9	9	9.20
1	R2	9	9	7	9	9	8.60
1	R3	9	9	8	9	9	8.80
1	Gold	9.00	9.00	8.33	9.00	9.00	8.87
2	R1	9	9	10	10	9	9.40
2	R2	8	8	8	9	9	8.40
2	R3	8	9	6	9	8	8.00
2	Gold	8.33	8.67	8.00	9.33	8.67	8.60
3	R1	8	9	10	10	8	9.00
3	R2	8	8	6	7	8	7.40
3	R3	9	8	9	7	8	8.20
3	Gold	8.33	8.33	8.33	8.00	8.00	8.20

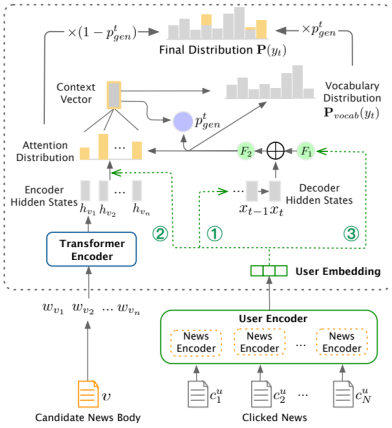
Case 2. Three figures. Figure 1 is a 2×3 grid of grouped bars, one panel per language code, with three series in each legend. The vertical axis is the share of novel n-grams. The caption states the metric, the contrast between neural outputs and human references, and that six datasets are shown. It does not name the legend entries or the six language codes. Figure 1 receives $(9, 9, 10, 9, 9)$,



(a) Figure 1. $\bar{y} = 8.27$.



(b) Figure 2. $\bar{y} = 8.07$.



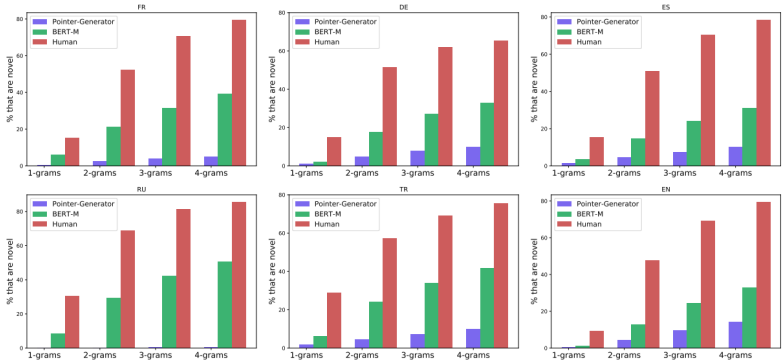
(c) Figure 3. $\bar{y} = 8.73$.

Figure A3: Case 1. Scores are in Table A15.

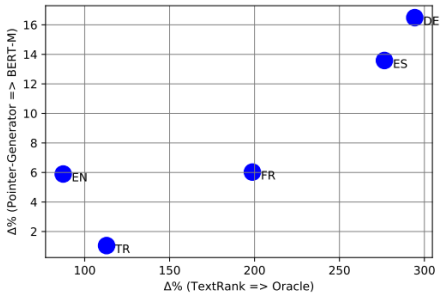
$(9, 9, 7, 9, 9)$, and $(9, 9, 8, 9, 9)$, so $\bar{s} = (9.00, 9.00, 8.33, 9.00, 9.00)$ and $\bar{y} = 8.87$. VC, SL, CTX, and MR are 9 for every rater. CC is $10/7/8$. The score of 7 is the rater who wants those names written into the caption. The citing paragraph discusses the same novelty comparison, including one language where copying dominates, so CTX does not follow the CC dip. SL spread is 0, in line with the tight SL column in Table A13.

Figure 2 is a scatter plot with five labeled points. Both axes are titled on the canvas, and the caption defines the same two improvement rates. Figure 2 receives $(9, 9, 10, 10, 9)$, $(8, 8, 8, 9, 9)$, and $(8, 9, 6, 9, 8)$, so $\bar{s} = (8.33, 8.67, 8.00, 9.33, 8.67)$ and $\bar{y} = 8.60$. CC is $10/8/6$. The score of 6 asks the caption to list the language codes already printed next to the points. CTX is $10/9/9$. The citing paragraph states the hypothesis the axes test, and it also says one language is left off the plot for scale. That missing point is not a silent omission, which is why CTX stays high while CC can still drop. MR is $9/9/8$.

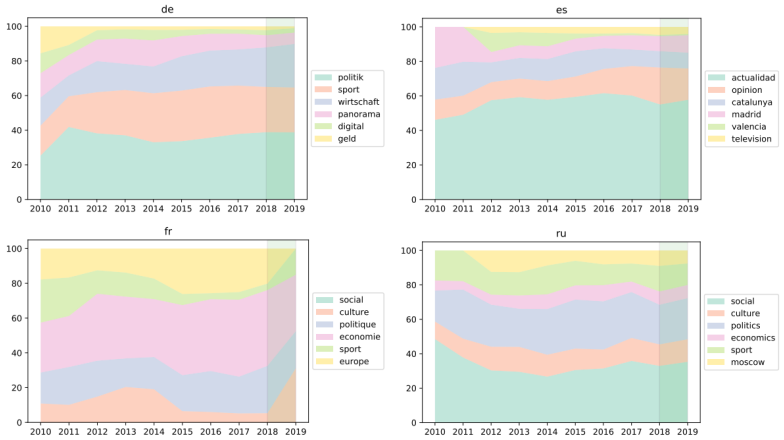
Figure 3 has four stacked-area panels, one language in each corner, with year on the horizontal axis. A shaded band marks 2019. The caption names the four panel positions and the meaning of the shade. Figure 3 receives $(8, 9, 10, 10, 8)$, $(8, 8, 6, 7, 8)$, and $(9, 8, 9, 7, 8)$, so $\bar{s} = (8.33, 8.33, 8.33, 8.00, 8.00)$ and $\bar{y} = 8.20$. CC is $10/6/9$. The score of 6 is because the topic names sit in the legends and not in the caption. Some neighboring colors are close, so VC is $8/8/9$ rather than 9 throughout, and MR is 8 for every rater. CTX is $10/7/7$. Two raters note that the citing paragraph describes the shift over categories, and the six most frequent categories, but does not mention the shaded band. The band is a caption fact, not a citing-text fact. The three figures keep overall gold between 8.20 and 8.87, while single dimensions still move by several points.



(a) Figure 1. $\bar{y} = 8.87$.



(b) Figure 2. $\bar{y} = 8.60$.



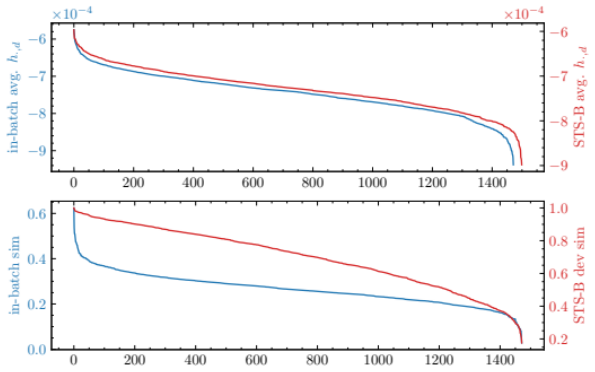
(c) Figure 3. $\bar{y} = 8.20$.

Figure A4: Case 2. Scores are in Table A16.

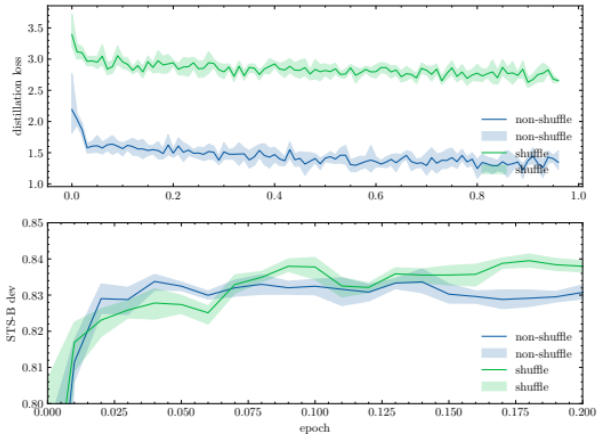
Case 3. Two figures. Figure 1 has two stacked panels and a shared horizontal axis with no label. Each panel has a left axis and a right axis, colored to match the two curves. The upper pair compares an in-batch average with a development-set average. The lower pair compares in-batch similarity with development-set similarity. The lines are sharp, which is why VC gold is 8.67. The caption names sharpness of logits for in-batch sentences and the development set, and it does not say which panel is the average and which panel is the similarity. Figure 1 receives $(9, 7, 6, 7, 7)$, $(8, 8, 7, 9, 8)$, and $(9, 8, 5, 8, 5)$, so $\bar{s} = (8.67, 7.67, 6.00, 8.00, 6.67)$ and $\bar{y} = 7.40$. CC is $6/7/5$. MR is $7/8/5$. The score of 5 on MR comes from the unlabeled horizontal axis together with axis titles ($h_{,d}$ and similarity) that the caption never defines. SL is $7/8/8$ because the panels have no subfigure letters. CTX is $7/9/8$. The citing paragraph discusses logit variance and similarity, so it supports the figure more than the caption does. This is the lowest overall gold in the five articles, and the drop is in CC and MR, not in VC.

Table A17: Case 3. Dimension scores and gated overall.

Fig.	Rater	VC	SL	CC	CTX	MR	Ov.
1	R1	9	7	6	7	7	7.20
1	R2	8	8	7	9	8	8.00
1	R3	9	8	5	8	5	7.00
1	Gold	8.67	7.67	6.00	8.00	6.67	7.40
2	R1	9	9	9	9	9	9.00
2	R2	8	7	9	9	8	8.20
2	R3	9	9	9	10	8	9.00
2	Gold	8.67	8.33	9.00	9.33	8.33	8.73



(a) Figure 1. $\bar{y} = 7.40$.



(b) Figure 2. $\bar{y} = 8.73$.

Figure A5: Case 3. Scores are in Table A17.

Figure 2 has two panels. The upper panel is a distillation-loss curve and the lower panel is a development score, each with a shuffle condition and a non-shuffle condition, plus a shaded band. The upper panel has no horizontal-axis title. One rater gives SL 7 for that. The caption names both curves and both strategies. Figure 2 receives $(9, 9, 9, 9, 9)$, $(8, 7, 9, 9, 8)$, and $(9, 9, 9, 10, 8)$, so $\bar{s} = (8.67, 8.33, 9.00, 9.33, 8.33)$ and $\bar{y} = 8.73$. CC is 9 for every rater. The citing paragraph describes the early drop of the non-shuffle loss and the continued drop under shuffling, which is what the two panels show, so CTX is $9/9/10$. The two figures use a similar line style. Overall gold differs by 1.33 because Figure 1 leaves the axes and the two metrics under-specified.

Table A18: Case 4. Dimension scores and gated overall.

Fig.	Rater	VC	SL	CC	CTX	MR	Ov.
1	R1	9	9	9	9	9	9.00
1	R2	9	9	9	9	9	9.00
1	R3	7	9	8	10	8	8.40
1	Gold	8.33	9.00	8.67	9.33	8.67	8.80
2	R1	8	9	8	7	9	8.20
2	R2	9	9	7	8	9	8.40
2	R3	9	9	9	6	6	7.80
2	Gold	8.67	9.00	8.00	7.00	8.00	8.13
3	R1	9	9	10	10	9	9.40
3	R2	8	9	9	9	9	8.80
3	R3	8	9	6	7	8	7.60
3	Gold	8.33	9.00	8.33	8.67	8.67	8.60

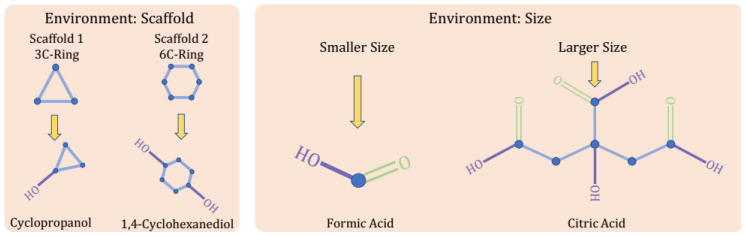
Case 4. Three figures. Figure 1 draws two examples side by side. The left pair keeps a small substituent and changes the ring scaffold. The right pair changes molecular size. Labels and arrows are complete, and the drawing is not cut off at the border. The caption states both examples, including which substructure is shared. Figure 1 receives $(9, 9, 9, 9, 9)$, $(9, 9, 9, 9, 9)$, and $(7, 9, 8, 10, 8)$, so $\bar{s} = (8.33, 9.00, 8.67, 9.33, 8.67)$ and $\bar{y} = 8.80$. Two raters score every dimension at 9. The third lowers VC to 7. CTX gold is 9.33, the highest dimension, because the citing paragraph walks through both examples in the same order as the drawing.

Figure 2 is a two-stage diagram. The upper band is labeled environment inference and the lower band is labeled representation learning, with a decompose arrow into substructures, attentive pooling, and a predictor. The caption describes that order and the two training losses. Figure 2 receives $(8, 9, 8, 7, 9)$, $(9, 9, 7, 8, 9)$, and $(9, 9, 9, 6, 6)$, so $\bar{s} = (8.67, 9.00, 8.00, 7.00, 8.00)$ and $\bar{y} = 8.13$. SL is 9 for every rater. Some internal labels are small, so one VC score is 8. CTX is $7/8/6$ and is the lowest gold dimension on this figure. The citing paragraph explains the substructure encoder and the attentive pooling in the lower band, and it does not walk through the upper band. One rater also scores MR at 6 for that gap. High SL with lower CTX is the split the rubric is meant to keep. Layout quality does not stand in for citing-text support.

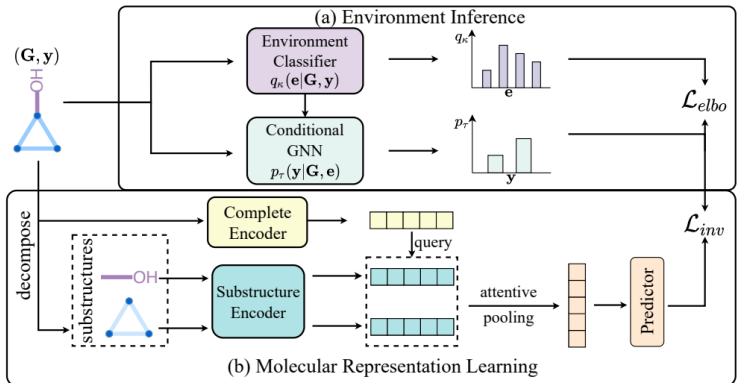
Figure 3 places two box plots beside a four-curve risk plot. Panel letters (a), (b), and (c) are printed under the axes. Panel (a) varies k, panel (b) varies β, and panel (c) shows risk against epoch for four environments. The vertical axis of (a) and (b) is ROC-AUC. Figure 3 receives $(9, 9, 10, 10, 9)$, $(8, 9, 9, 9, 9)$, and $(8, 9, 6, 7, 8)$, so $\bar{s} = (8.33, 9.00, 8.33, 8.67, 8.67)$ and $\bar{y} = 8.60$. SL is 9 for every rater. CC is $10/9/6$. The score of 6 is because the caption names k, β, and the risk curves, but does not name ROC-AUC. CTX is $10/9/7$. The citing paragraph discusses the trends in (a) and (c) in detail, and it mentions β for (b) without describing that panel's shape. Across the article, SL gold is 9.00 on every figure, while CTX gold moves from 9.33 to 7.00 and back to 8.67.

Case 5. Four figures. Figures 1, 3, and 4 are attention heat maps. Figure 2 is a grouped bar chart. Token axes and color bars are present, and the panels are not cropped. SL gold is 9.00 on every figure, with every rater at 9. Several VC scores are 7 because color-bar ticks and axis text are small. Layout stays stable, and the disagreement sits in CC, CTX, and MR. That repeats the *eval1200* pattern in Table A13.

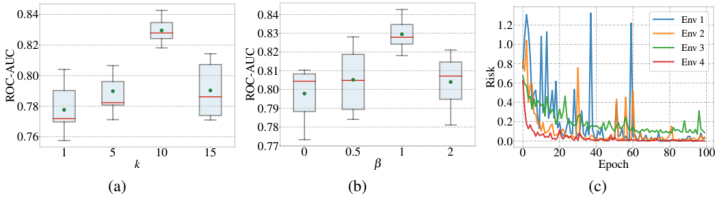
Figure 1 shows two maps side by side, titled $\sqrt{64}$ and $\sqrt{96}$. High weights form short diagonal bands, and on the right map those bands reach later token rows. The caption calls the left map normal and the right map smoothed, and it does not tie those words to the two temperature titles. Figure 1 receives $(8, 9, 8, 8, 9)$, $(8, 9, 8, 9, 9)$, and $(7, 9, 4, 9, 7)$, so $\bar{s} = (7.67, 9.00, 6.67, 8.67, 8.33)$ and $\bar{y} = 8.07$. CC is $8/8/4$. The score of 4 is that missing link between the caption wording and the panel titles. The same rater gives MR 7 for the same reason, and CTX 9, because the citing paragraph describes shorter copy spans and attention that extends past the first 200 tokens. Gold CC is the dimension-first mean of 8, 8, and 4, the same aggregation as the worked example in Section D.3. CC spread on this figure is 4, above the *eval1200* median CC spread of 3.0.



(a) Figure 1. $\bar{y} = 8.80$.



(b) Figure 2. $\bar{y} = 8.13$.



(c) Figure 3. $\bar{y} = 8.60$.

Figure A6: Case 4. Scores are in Table A18.

Figure 2 plots the share of attention weights at or above 0.15, in five bins of normalized token position. The legend is λ=1.0, 1.5, and 2.0. The caption says the comparison is across attention temperatures, and it does not mention λ. Figure 2 receives $(9, 9, 9, 9, 9)$, $(9, 9, 9, 10, 9)$, and $(9, 9, 4, 7, 9)$, so $\bar{s} = (9.00, 9.00, 7.33, 8.67, 9.00)$ and $\bar{y} = 8.60$. VC, SL, and MR are 9 for every rater. CC is again $9/9/4$, from the temperature wording against the λ legend. CTX is $9/10/7$. The citing paragraph defines the 0.15 cutoff, the five bins, and the shift from the head of the document toward the tail, so two raters score CTX at 9 and 10 even though one of them scores CC at 4. Gold CC 7.33 is $(9{+}9{+}4)/3$. One harsh caption reading moves the gold, and it does not move SL.

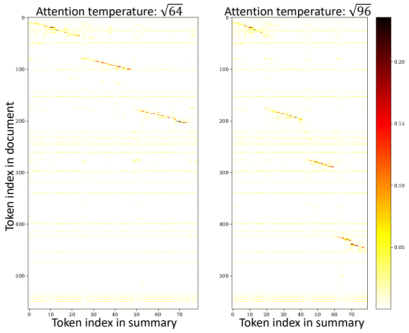
Figure 3 shows three maps in one row, at $\sqrt{64}$, $\sqrt{96}$, and $\sqrt{128}$. The color bar runs higher than in Figure 1, up to about 0.6, and a dark diagonal is visible in the left and middle panels. Figure 3 receives $(7, 9, 8, 9, 9)$, $(9, 9, 7, 8, 8)$, and $(9, 9, 8, 4, 7)$, so $\bar{s} = (8.33, 9.00, 7.67, 7.00, 8.00)$ and $\bar{y} = 8.00$. CC is $8/7/8$. The middle score wants the three temperature values written into the caption. The split that matters is CTX, at $9/8/4$. The citing paragraph discusses a shorter summary and a phrase taken from the end of the document. It does not map those outputs onto $\sqrt{64}$, $\sqrt{96}$, and $\sqrt{128}$, and the heat maps show token indices rather than the words themselves. That is why one rater scores CTX at 4 and MR at 7, while the other two still find the paragraph supportive.

Figure 4 is a second three-panel example in the same format. The document axis runs past 700 tokens, unlike Figure 3, whose axis stops near 200. The left and middle maps stay near the top of that axis. The right map also marks bands near 350 and near 700. Figure 4 receives $(7, 9, 8, 9, 9)$, $(9, 9, 6, 9, 7)$, and $(9, 9, 7, 7, 7)$, so $\bar{s} = (8.33, 9.00, 7.00, 8.33, 7.67)$ and $\bar{y} = 8.07$. CC is $8/6/7$

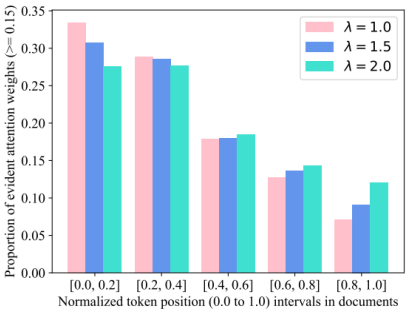
Table A19: Case 5. Dimension scores and gated overall.

Fig.	Rater	VC	SL	CC	CTX	MR	Ov.
1	R1	8	9	8	8	9	8.40
1	R2	8	9	8	9	9	8.60
1	R3	7	9	4	9	7	7.20
1	Gold	7.67	9.00	6.67	8.67	8.33	8.07
2	R1	9	9	9	9	9	9.00
2	R2	9	9	9	10	9	9.20
2	R3	9	9	4	7	9	7.60
2	Gold	9.00	9.00	7.33	8.67	9.00	8.60
3	R1	7	9	8	9	9	8.40
3	R2	9	9	7	8	8	8.20
3	R3	9	9	8	4	7	7.40
3	Gold	8.33	9.00	7.67	7.00	8.00	8.00
4	R1	7	9	8	9	9	8.40
4	R2	9	9	6	9	7	8.00
4	R3	9	9	7	7	7	7.80
4	Gold	8.33	9.00	7.00	8.33	7.67	8.07

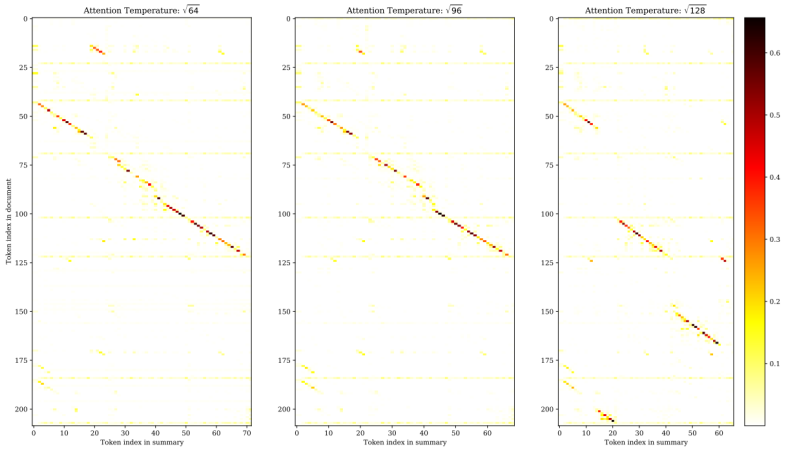
because the caption calls this a second example and does not say that the document is long, which is the visible difference from Figure 3. MR is $9/7/7$ for the same omission. CTX is $9/9/7$. The citing paragraph states that the document has more than 700 words and that one setting stays near the first 150 words while another also uses the middle and the tail. Here the citing text carries the claim that the caption leaves out. Overall gold on the four figures is 8.07, 8.60, 8.00, and 8.07. The article stays near the *eval1200* mean. The movement is inside CC, CTX, and MR, not in SL.



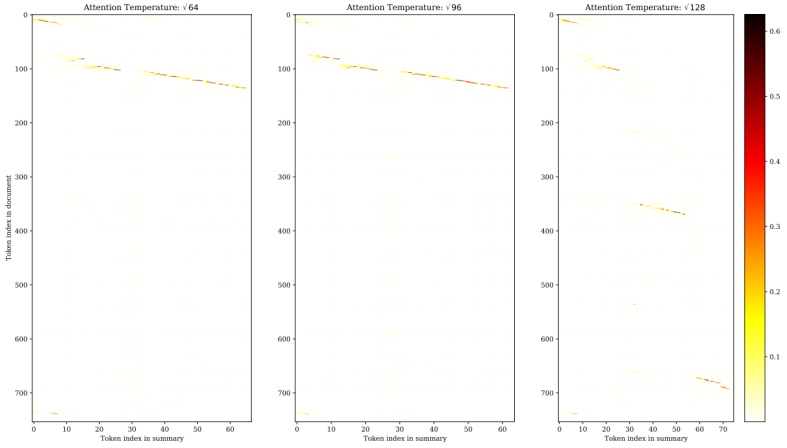
(a) Figure 1. $\bar{y} = 8.07$.



(b) Figure 2. $\bar{y} = 8.60$.



(c) Figure 3. $\bar{y} = 8.00$.



(d) Figure 4. $\bar{y} = 8.07$.

Figure A7: Case 5. Scores are in Table A19.