VISEXMEM: VISUAL EVIDENCE-BASED EXPLAINABLE MULTIMODAL EVALUATION METRIC

Anonymous authors
Paper under double-blind review

ABSTRACT

Current image–text evaluation often relies on a global score, which gives limited insight into which semantic details are supported by the image. Explainable metrics provide rationales, but the visual evidence supporting those rationales can still be difficult to verify. We introduce VISEXMEM, a reference-free evaluator that decomposes captions into atomic claims, localizes visual evidence, verifies each claim, and averages claim-level support into a global score. Explanations use the same evidence and decisions. We train Qwen3.5-based 4B and 9B models on 100K samples with multi-task SFT, counterfactual refinement, and evidence-aware GDPO. Across eight benchmarks, VISEXMEM-9B achieves the highest descriptive composite among open-source evaluators in our comparison, exceeding VQAScore by 11.7 points while remaining competitive with GPT-5.6 Sol. We also introduce AGITA-BENCH, containing 600 human-annotated image–text pairs and 8.7K atomic claims for global and claim-level evaluation. For untuned Qwen3.5 and GPT-5.6 Sol controls, decomposed prompting raises the same composite by 3.6–5.1 points relative to holistic prompting. Human evaluation shows that VISEXMEM explanations outperform prior explainable metrics and approach GPT-5.6 Sol.

1 INTRODUCTION

The rapid progress of multimodal generative models has made reliable image–text evaluation increasingly important for benchmarking and model development (Hartwig et al., 2025; Berger et al., 2025; Sarto et al., 2025). Automatic evaluation has progressed from global cross-modal similarity (Hessel et al., 2021) to learned preference models (Kirstain et al., 2023; Xu et al., 2023; Wu et al., 2023), structured VQA-based evaluation (Hu et al., 2023; Cho et al., 2024; Lin et al., 2024), VLM-as-a-judge approaches (Chen et al., 2025; Pu et al., 2025; Li et al., 2026), and structured alignment functions (Chen et al., 2026a; Koyama et al., 2026). However, a high global score can still hide a local semantic error. An image may look broadly correct while getting a count, attribute binding, action, spatial relation, or subject–object role wrong. Benchmarks such as Winoground, TIFA, GenAI-Bench, MMIG-Bench, and ImagenWorld are designed to test these fine-grained and compositional errors (Thrush et al., 2022; Hu et al., 2023; Li et al., 2024a; Hua et al., 2025; Mahdizadeh Sani et al., 2026).

Strong evaluation performance also does not make a metric easy to inspect. VIEScore and FLEUR provide natural-language rationales (Ku et al., 2024; Lee et al., 2024), while recent work adds criterion-level explanations (Kim et al., 2025), multi-criteria aggregation (Oi et al., 2026), and element-level grounding (Shi et al., 2026). A remaining issue is whether the explanation reflects the visual evidence and the decisions that produced the score. A fluent explanation can still sound reasonable even when it is based on the wrong evidence or decision. Similarly, global score correlation does not show whether a metric responds to the specific visual or linguistic constraint that is wrong (Ross et al., 2024).

We introduce VISEXMEM (***Vis**ual Evidence-Based **Ex**plainable **M**ultimodal **E**valuation **M**etric*), a reference-free metric that decomposes text into atomic claims and localizes visual evidence (Figure 1). For each claim, the localized evidence is used to compute a Yes/No log-probability margin, which is mapped to a claim support score. The arithmetic mean of these scores gives the global score. Explanations are generated afterward from the same evidence and fixed claim decisions, so

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.

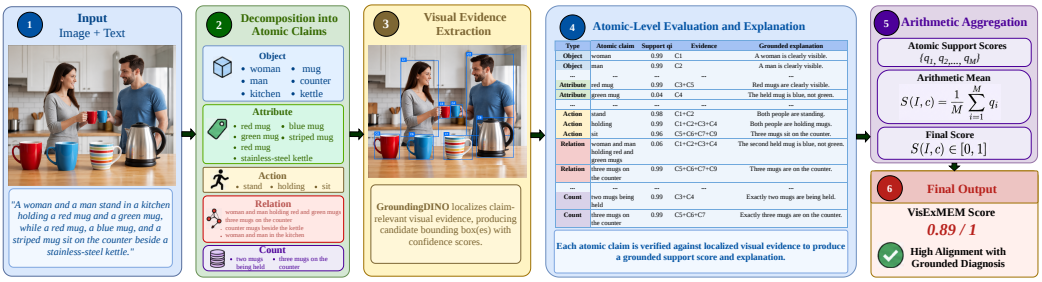

Figure 1: **Overview of VisExMEM.** VisExMEM decomposes captions into atomic claims, localizes visual evidence for each claim, computes claim support scores, and averages them into a global score. Explanations are generated afterward from the same evidence and fixed claim decisions.

they do not determine the score. We train Qwen3.5-based (Qwen Team, 2026) 4B and 9B models in three stages (Figure 2). Multi-task SFT establishes the evaluation pipeline and its structured outputs. Counterfactual refinement uses minimal semantic changes so the model responds to a single changed constraint while keeping the remaining content fixed. Finally, evidence-aware GDPO (Liu et al., 2026) improves semantic verification and encourages predictions to rely on localized visual evidence.

To evaluate both global and claim-level agreement with humans, we introduce AGITA-BENCH (***A**tomic and **G**lobal **I**mage–**T**ext **A**lignment Benchmark*), a 600-case benchmark with human annotations covering image-to-text (I2T) and text-to-image (T2I) evaluation. In I2T, captions are evaluated against real images. In T2I, generation prompts are evaluated against generated images. Each case belongs to an easy, medium, or hard compositional setting. Annotators provide an overall image–text consistency rating and claim-level support judgments. This allows us to evaluate both global and claim-level scores. Our contributions are:

- **Evidence-Grounded Explainable Evaluation:** We introduce VISEXMEM, which links atomic claim scores to localized visual evidence and generates explanations from the same evidence and claim decisions.
- **Progressive Multimodal Verification Training:** We develop a three-stage training strategy for learning structured evaluation, sensitivity to minimal semantic changes, and dependence on visual evidence.
- **AGITA-Bench:** We provide a benchmark that evaluates global agreement and individual claim decisions in both I2T and T2I settings across different levels of compositional difficulty.

2 RELATED WORK

Image–text alignment evaluation. Reference-free metrics include embedding-based methods that score an image–text pair as a whole, without decomposing it into smaller semantic units (Hessel et al., 2021; Li et al., 2023) and learned human-preference models (Kirstain et al., 2023; Xu et al., 2023; Wu et al., 2023; Zhang et al., 2024). More recent methods break the evaluation into smaller semantic units. TIFA (Hu et al., 2023) and DSG (Cho et al., 2024) decompose prompts into verifiable questions, VQAScore (Lin et al., 2024) uses generative VQA likelihood, and T2I-FineEval (Hosseini et al., 2025) decomposes both text and image. Recent approaches also explore hyperbolic alignment (Chen et al., 2026a) and task-adapted caption scoring (Koyama et al., 2026). VISEXMEM follows this fine-grained setting by decomposing the caption into atomic claims for evaluation.

Explainable and evidence-grounded evaluation. FLEUR (Lee et al., 2024) and VIEScore (Ku et al., 2024) pair scores with natural-language rationales, while EXPERT (Kim et al., 2025) and HarmonicEval (Oi et al., 2026) expose structured criteria. Grounding-oriented work includes GLaMM (Rasheed et al., 2024), Grounding-IQA (Chen et al., 2026b), and multimodal-rationale VQA (Li et al., 2025). Most closely, REVEALER (Shi et al., 2026) performs element-level align-

ment evaluation through a grounding–reasoning–scoring trajectory. VISEXMEM instead uses localized visual evidence to score each atomic claim. These claim scores determine the global score, while the explanation is generated afterward from the same evidence and fixed claim decisions.

Diagnostic benchmarks and metric meta-evaluation. Benchmarks test different aspects of image–text evaluation. Winoground (Thrush et al., 2022) and EqBen (Wang et al., 2023a) probe compositional sensitivity; TIFA (Hu et al., 2023), GenAI-Bench (Li et al., 2024a), T2I-CompBench++ (Huang et al., 2025), and CountBench (Paiss et al., 2023) assess fine-grained faithfulness, compositionality, and counting, while Pick-a-Pic (Kirstain et al., 2023) provides human preference comparisons. Broader diagnostic resources include EvalMuse-40K (Han et al., 2026), MMIG-Bench (Hua et al., 2025), FineGRAIN (Hayes et al., 2025), and Imagen-World (Mahdizadeh Sani et al., 2026). For metric meta-evaluation, PROTOBIAS (Roy et al., 2026) tests prototypicality bias, T2IScoreScore (Saxon et al., 2024) evaluates ordering and separation over semantic error graphs, and CROC (Leiter et al., 2026) uses pseudo- and human-labeled contrastive robustness checks. AGITA-BENCH complements these benchmarks by pairing a global rating with claim-level support judgments for the same image–text case, allowing overall agreement and individual claim decisions to be evaluated together.

3 VISEXMEM

VISEXMEM is a reference-free image–caption evaluation metric that grounds claim-level verification in localized visual evidence. In this section, we describe its scoring formulation, training procedure, and grounded explanation mechanism.

3.1 EVIDENCE-GROUNDED VERIFICATION AND SCORING

Given image I and caption c, VISEXMEM decomposes c into M typed atomic claims $\mathcal{A}(c) = \{(a_i, t_i)\}_{i=1}^{M}$, where a_i is the claim and t_i its type (*object, attribute, action, relation, count or negation*). GroundingDINO (Liu et al., 2024) localizes claim-relevant regions as visual evidence. As shown in Figure 1, the model evaluates each claim using this evidence with the full image to retain scene context.

For each claim, the model assigns mean token log-probabilities s_i^Y and s_i^N to `yes` and `no`. We define $m_i = s_i^Y - s_i^N$ and $q_i = \sigma(m_i)$, where $\sigma(x) = 1/(1 + e^{-x})$. Applying sigmoid after subtraction maps the Yes/No margin to $[0, 1]$; applying it first would compare compressed scores. The global score is $S(I, c) = M^{-1} \sum_{i=1}^{M} q_i$. We retain each claim's type, support score, localized box(es), grounding confidence, and decision. Explanations are generated afterward from these fixed records and cannot alter the score.

3.2 TRAINING

We train Qwen3.5-4B and Qwen3.5-9B (Qwen Team, 2026) VISEXMEM models with our novel three-stage training procedure (Figure 2): (1) multi-task SFT directly supervises decomposition, claim verification, evidence selection, and explanation, (2) counterfactual refinement teaches the model to distinguish minimally incorrect semantic edits from supported claims, and (3) evidence-aware GDPO (Liu et al., 2026) optimizes claim verification to depend on localized visual evidence.

Training data. Training data come from MS COCO (Lin et al., 2014), Visual Genome (Krishna et al., 2017), VSR (Liu et al., 2023), Flickr30k Entities (Plummer et al., 2015), Open Images (Kuznetsova et al., 2020) with Localized Narratives (Pont-Tuset et al., 2020), DOCCI (Onoe et al., 2024), PixMo-Count (Deitke et al., 2025), CROC_{syn} (Leiter et al., 2026), and the training split of EvalMuse-40K (Han et al., 2026). We convert the text into typed atomic claims, using native annotations and human judgments when available and an offline Qwen3.6-27B teacher for missing decomposition or structure; uncertain outputs are filtered out. GroundingDINO then localizes candidate boxes as visual evidence cues, which are provided as evidence rather than SFT targets. We use boxes instead of pixel masks because they localize the relevant region without revealing its exact silhouette, reducing mask-shape shortcuts. We construct 40K SFT examples, a 20K counterfactual source pool, and 40K Stage 3 examples equally split between semantic-verification and

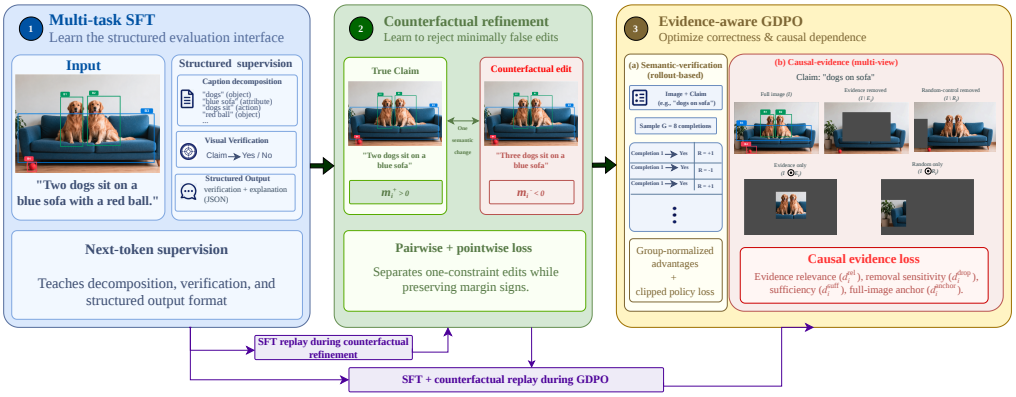

Figure 2: **VisExMEM training pipeline.** Multi-task SFT learns the structured evaluation tasks, counterfactual refinement targets semantic changes, and evidence-aware GDPO improves verification and evidence dependence. Earlier supervision is replayed in later stages.

causal-evidence supervision (see Appendix Figure 7). Counterfactual data combine CROC_{syn} image pairs with text-side minimal pairs that change one verified semantic constraint while keeping the image fixed. Explanations are constructed only after claims, decisions, and evidence are fixed. Full construction, filtering, source composition, and replay schedules are in Appendix A.1.

Stage 1: Multi-task SFT. Stage 1 teaches the structured evaluation tasks used by VisExMEM. We jointly supervise caption decomposition, claim-level verification, evidence selection, and explanation generation through next-token supervision. For image I, caption c, and structured target $y = (y_1, \ldots, y_T)$, we minimize

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E} \sum_{t=1}^{T} \log \pi_\theta \left(y_t \mid y_{<t}, I, c \right). \tag{1}$$

where π_θ is the trainable model. Candidate boxes are supplied as visual evidence; evidence selection predicts candidate IDs rather than box coordinates. Representative processed records are shown in Appendix Figure 7; exact target formats, task mixtures, and optimization settings are given in Appendix A.2.3.

Stage 2: Counterfactual refinement. Stage 2 makes claim verification sensitive to controlled semantic changes. Each supported image–claim pair is paired with a contradicted counterpart. Text-side pairs change one constraint while keeping the image fixed, e.g., "two dogs" to "three dogs," while image-side pairs provide the corresponding visual contrast. Let

$$m_\theta(I, a) = \ell_\theta(\texttt{yes} \mid I, a) - \ell_\theta(\texttt{no} \mid I, a), \tag{2}$$

where ℓ_θ denotes the decision-token logit under model parameters θ. Training requires a positive margin for the supported example, a negative margin for the counterfactual, and their difference to exceed a target margin. This is supervised margin training, not reinforcement learning. Stage 1 decomposition and verification examples are replayed to retain earlier supervision (Luo et al., 2026); the full loss, filtering rules, and replay settings are given in Appendix A.2.4.

Stage 3: Evidence-aware GDPO. Stage 3 improves semantic verification while making predictions depend on localized visual evidence. Building on GRPO-style group-relative optimization (Shao et al., 2024), we use an evidence-aware GDPO variant inspired by GDPO (Liu et al., 2026): semantic correctness provides the group-relative policy signal, while evidence dependence is optimized through a separate causal objective. For semantic input x_i, we sample G completions $o_{i,g} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x_i)$, where g indexes rollouts and $\pi_{\theta_{\text{old}}}$ is the rollout policy. For semantic reward $r_{i,g}$, let μ_i and σ_i be the group mean and standard deviation, and define $h_{i,g,t} = (x_i, o_{i,g,<t})$. The normalized advantage and policy ratio are

$$A_{i,g} = \frac{r_{i,g} - \mu_i}{\sigma_i}, \qquad \rho_{i,g,t} = \frac{\pi_\theta(o_{i,g,t} \mid h_{i,g,t})}{\pi_{\theta_{\text{old}}}(o_{i,g,t} \mid h_{i,g,t})}. \tag{3}$$

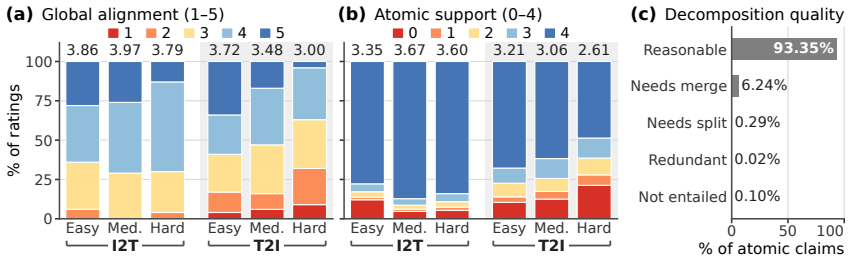

Figure 3: **AGITA-Bench annotation summary.** (a) Global-alignment and (b) atomic-support distributions by direction and difficulty; means are shown above bars. (c) Decomposition quality.

Here, $A_{i,g}$ compares each rollout with its group, while $\rho_{i,g,t}$ measures the token-level change from the rollout policy to the updated policy. Numerical stabilization and zero-variance handling are given in Appendix A.2.5. Following the clipped GRPO objective, we optimize

$$\mathcal{J}_{\text{sem}} = \mathbb{E}\left[\frac{1}{G}\sum_{g=1}^{G}\frac{1}{|o_{i,g}|}\sum_{t}\left(\min\{\rho_{i,g,t}A_{i,g}, \text{clip}(\rho_{i,g,t}, 1-\epsilon, 1+\epsilon)A_{i,g}\} - \beta d_{i,g,t}^{\text{KL}}\right)\right], \quad (4)$$

where ϵ is the clipping radius, β weights KL regularization, and $d_{i,g,t}^{\text{KL}}$ penalizes deviation from the frozen Stage 2 reference.

Semantic reward alone does not ensure evidence dependence. For each eligible supported claim, selected evidence E_i is paired with a matched random region R_i, and $s_i(V)$ denotes the verification margin under image view V. We encourage $s_i(I) - s_i(I \setminus E_i) > s_i(I) - s_i(I \setminus R_i)$ and $s_i(I \odot E_i) > s_i(I \odot R_i)$, so selected evidence is more necessary and sufficient than its control; $I \setminus E$ masks region E, while $I \odot E$ retains only it. Stage 3 minimizes $\mathcal{L}_{\text{Stage3}} = -\mathcal{J}_{\text{sem}} + \lambda_c \mathcal{L}_{\text{causal}}$, where λ_c weights the causal loss, using a 7:2:1 GDPO:counterfactual:SFT update mixture. Further details are in Appendix A.2.5; the Verification-GRPO and Causal-GRPO ablations branch from Stage 2 and are described in Appendix A.3.

4 AGITA-BENCH

To evaluate both global and claim-level agreement with humans, we introduce AGITA-BENCH (***A**tomic and **G**lobal **I**mage–**T**ext **A**lignment*), a 600-pair benchmark with human judgments at both levels. It contains 300 I2T real photographs, with 100 each from ADE20K (Zhou et al., 2017), HICO-DET (Chao et al., 2018), and Objects365 (Shao et al., 2019), and 300 T2I pairs from DiffusionDB (181) (Wang et al., 2023b) and MJHQ-30K (119) (Li et al., 2024b). I2T captions cover high, partial, and low alignment, while T2I uses the original generation prompts. We balance 200 easy, 200 medium, and 200 hard cases by *semantic complexity*, not alignment severity. GPT-5.6 Sol decomposes the text into atomic claims, and fixed rules assign difficulty from claim number and composition, easy cases contain few simple claims, medium cases add relations or attribute/count assignments, and hard cases require more constraints or correct binding of attributes, counts, roles, or spatial relations. This produces 8,735 claims spanning objects, attributes, actions, relations (including spatial relations), counts, and negation. Representative examples are shown in Figure 4; AGITA-Bench is excluded from training, calibration, and model selection.

Annotators rate atomic support from 0 (contradicted/not supported) to 4 (fully supported) and global alignment from 1 (not aligned) to 5 (fully aligned), with *cannot judge* excluded from atomic statistics. Decomposition quality is rated separately; *reasonable as written* means a clear, non-redundant atomic claim requiring neither splitting nor merging, independent of image support. Mean global alignment is 3.64/5 and atomic support 3.35/4, while 93.35% of claims are judged reasonable and 2.07% receive *cannot judge*. Figure 3 shows that I2T ratings remain relatively high across complexity levels, whereas T2I shifts toward lower global and atomic ratings as complexity increases. All seven annotators independently rate a shared set of 30 pairs (460 unique claims), giving mean pairwise quadratic-weighted Cohen's κ of .602 for global alignment and .564 for atomic support; the remaining 570 pairs are rated once. Full annotation instructions, agreement results, annotator details, and the decomposition prompt are given in Appendix A.6.



Figure 4: **Qualitative AGITA-Bench examples with VISEXMEM-9B.** Easy–Hard cases show the input, localized evidence, caption/prompt, and Human/VISEXMEM-9B global and atomic scores (H/V); human ratings are normalized to $[0, 1]$, with short claim-level explanations.

5 EXPERIMENTS

Benchmarks and Protocols. We evaluate ten benchmarks, comprising eight established benchmarks, T2IScoreScore (TS2), and AGITA-Bench. We report Group scores on Winoground (WG) (Thrush et al., 2022) and EqBen (Wang et al., 2023a), pairwise accuracy (PAcc) on TIFA160 (Hu et al., 2023) and GenAI-Bench (Li et al., 2024a), and preference accuracy (Pref. Acc.) on Pick-a-Pic (Kirstain et al., 2023). For $\mathrm{CROC}_{\mathrm{hum}}$ (Leiter et al., 2026), we use the mean over its forward and inverse text-based and image-based evaluations. We report SRCC on the T2I-CompBench++ (T2I-CB++) human subset (Huang et al., 2025) and top-1 counting accuracy on CountBench (Paiss et al., 2023). TS2 (Saxon et al., 2024) organizes images into semantic error graphs (SEGs). We report ordering rank_m and separation sep_m over all SEGs and the Synth, Nat, and Real subsets. Synth uses generated images for predefined error graphs, Nat uses natural images in constructed graphs, and Real uses naturally occurring T2I errors. AGITA-Bench (Sec. 4) reports Spearman's ρ for global and atomic human agreement across Easy, Medium, and Hard. Full protocols are in Appendix A.5. We also evaluate OOD generalization on VisPlotBench (Ni et al., 2026), CulturalFrames (Nayak et al., 2025), and ProtoBias (Roy et al., 2026) in Appendix A.4.1.

Evaluation Metrics. We compare VISEXMEM-4B and VISEXMEM-9B with **embedding- and matching-based** CLIPScore (Hessel et al., 2021), BLIPv2Score (Li et al., 2023), and ALIGNScore (Saxon et al., 2024), **preference-trained** PickScore (Kirstain et al., 2023), ImageReward (Xu et al., 2023), HPSv2 (Wu et al., 2023), and MPS (Zhang et al., 2024), **general-purpose or explainable evaluators** GPT4V-Eval (Zhang et al., 2023), VIEScore (Ku et al., 2024), FLEUR, and EXPERT, and **structured and VQA-based** VQ^2 (Yarom et al., 2023), Davidsonian Scene Graph (Cho et al., 2024), our DSG + GPT-5.6 Sol setup, and VQAScore (Lin et al., 2024). We also evaluate GroundingDINO (Liu et al., 2024) directly as a grounding-based metric. Controlled **VLM-as-a-Judge** experiments use GPT-5.6 Sol and untuned Qwen3.5 (Qwen Team, 2026) with *Holistic (Hol.)* prompting on the full image-caption pair or *Decomposed (Dec.)* prompting with a decompose, verify, and aggregate procedure, without VISEXMEM-specific training, learned evidence conditioning, or external grounding cues. On AGITA-Bench, global agreement uses native scores, while atomic agreement is reported only for methods directly scoring the fixed benchmark claims. Prompts are

Table 1: **Main results.** Comparison of prior image–text alignment metrics, controlled VLM-as-a-Judge baselines, and VISEXMEM across eight established benchmarks. Higher is better. **Bold**, underlined, and *italicized* values mark the best, second-best, and third-best results.

Method	WG Group	EqBen Group	TIFA160 PAcc	Pick-a-Pic Pref. Acc.	GenAI PAcc	CROC_{hum} Score	T2I-CB++ SRCC	CountBench Acc.	Avg.
CLIPScore	7.8	25.0	54.1	76.0	52.2	69.8	28.5	30.7	43.0
BLIPv2Score	17.5	35.0	57.5	80.0	55.1	76.1	37.0	47.3	50.7
ALIGNScore	8.3	28.6	59.6	80.0	53.4	71.8	36.1	34.5	46.5
PickScore	6.8	23.6	59.4	70.0	57.7	74.7	30.3	56.5	47.4
ImageReward	12.8	26.4	67.3	75.0	58.7	75.4	50.6	33.8	50.0
HPSv2	4.0	17.1	55.2	69.0	49.8	72.3	42.4	*66.6*	47.0
MPS	1.8	10.0	59.0	69.0	56.1	67.0	24.2	33.8	40.1
GPT4V-Eval	36.3	35.0	64.0	74.0	61.4	–	–	–	54.1†
VIEScore	34.5	32.9	63.9	78.0	60.3	86.5	54.9	47.3	57.3
EXPERT	29.3	27.1	68.0	83.0	64.6	83.0	54.2	43.9	56.6
FLEUR	28.5	35.7	67.2	82.0	63.5	81.7	52.2	41.0	56.5
VQ^2	10.0	20.0	48.7	73.0	51.9	78.1	48.6	51.1	47.7
Davidsonian	15.5	20.0	54.3	70.0	54.6	77.7	*60.6*	37.9	48.8
DSG + GPT-5.6 Sol	49.0	40.6	66.8	81.0	61.4	72.1	49.8	51.3	59.0
VQAScore	46.0	47.9	71.2	*84.0*	63.3	86.1	53.2	31.7	60.4
GroundingDINO	9.3	12.9	58.2	57.0	51.1	56.8	18.5	20.6	35.6
Qwen3.5-4B (Hol.)	28.2	44.0	64.3	72.5	62.2	78.1	56.7	36.6	55.3
Qwen3.5-4B (Dec.)	49.0	47.1	66.0	80.0	62.5	76.5	57.2	44.2	60.3
Δ *Dec.–Hol.*	(+20.8)	(+3.1)	(+1.7)	(+7.5)	(+0.3)	(-1.6)	(+0.5)	(+7.6)	(+5.0)
Qwen3.5-9B (Hol.)	40.4	35.7	50.1	74.2	50.3	76.2	55.8	25.5	51.0
Qwen3.5-9B (Dec.)	41.2	45.7	62.2	81.0	54.6	74.4	54.8	35.0	56.1
Δ *Dec.–Hol.*	(+0.8)	(+10.0)	(+12.1)	(+6.8)	(+4.3)	(-1.8)	(-1.0)	(+9.5)	(+5.1)
GPT-5.6 Sol (Hol.)	58.5	**70.1**	*71.8*	82.4	*66.9*	*91.0*	**63.8**	**87.6**	74.0
GPT-5.6 Sol (Dec.)	**80.0**	70.0	**74.3**	**86.0**	**70.5**	**94.0**	61.1	85.1	**77.6**
Δ *Dec.–Hol.*	(+21.5)	(-0.1)	(+2.5)	(+3.6)	(+3.6)	(+3.0)	(-2.7)	(-2.5)	(+3.6)
VisExMEM-4B (Ours)	*65.6*	*59.1*	70.2	80.2	66.3	87.0	57.8	51.9	67.3
vs. VQAScore	(+19.6)	(+11.2)	(-1.0)	(-3.8)	(+3.0)	(+0.9)	(+4.6)	(+20.2)	(+6.8)
VisExMEM-9B (Ours)	68.9	**70.1**	73.1	85.6	69.4	91.7	57.7	60.7	*72.2*
vs. VQAScore	(+22.9)	(+22.2)	(+1.9)	(+1.6)	(+6.1)	(+5.6)	(+4.5)	(+29.0)	(+11.7)

† GPT4V-Eval's Avg. uses its five available benchmark results; all other Avg. values use all eight. T2I-CB++ reports 100ρ (theoretical range $[-100, 100]$). Avg. is a descriptive composite of different benchmark statistics. Blue rows show signed differences relative to VQAScore in the displayed units; T2I-CB++ differences are $100\Delta\rho$, not percentage points.

given in Appendix A.5.1. The omission of REVEALER (Shi et al., 2026) and T2I-FineEval (Hosseini et al., 2025) is discussed in Appendix A.5.2.

Explanation Evaluation. We compare EXPERT, VIEScore, GPT-5.6 Sol (Dec.), and VISEXMEM-9B. Annotators rate each explanation from 1 to 5 for diagnostic correctness, visual grounding, clarity, and diagnostic usefulness. Correctness measures whether the explanation identifies the relevant matches or errors, grounding whether its statements are supported by the image and caption, clarity how easy it is to understand, and usefulness how well it explains why the image and caption agree or disagree. Annotators also flag unsupported factual content. For systems with localized evidence, they rate localization accuracy and sufficiency from 1 to 5. Localization measures whether the highlighted region identifies the relevant visual evidence, while sufficiency measures whether that evidence is enough to judge the claim. Full instructions and agreement statistics are given in Appendix A.7.

6 RESULTS & ANALYSIS

Comparison with Existing Metrics. Table 1 compares performance across the eight established benchmarks. VISEXMEM-9B has the highest composite among open-source evaluators in our comparison. It exceeds VQAScore on all eight, with an 11.7-point composite difference. VISEXMEM-4B improves the composite by 6.8 points over VQAScore and exceeds all prior open-source baselines. For 9B, the largest gains are on Winoground (+22.9), EqBen (+22.2), and CountBench (+29.0), with a further +6.1 points on GenAI-Bench. VISEXMEM-9B also exceeds GPT-5.6 Sol (Hol.) on five benchmarks and matches it on EqBen. Its eight-benchmark composite is 72.2, compared with 74.0 for GPT-5.6 Sol (Hol.) and 77.6 for GPT-5.6 Sol (Dec.).

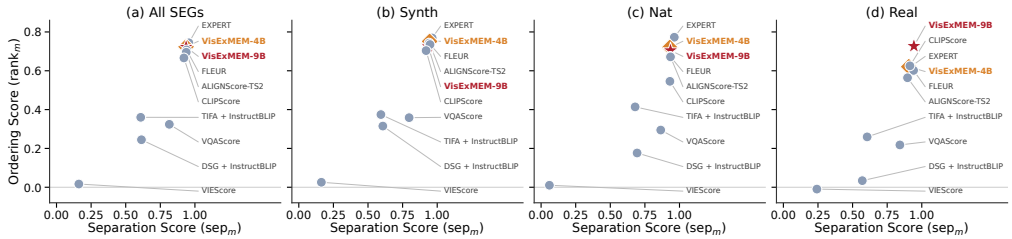

Figure 5: **T2IScoreScore performance.** Ordering (rank_m) versus separation (sep_m) for (a) all SEGs, (b) Synth, (c) Nat, and (d) Real. Higher is better on both axes.

Table 2: **Human Agreement on AGITA-Bench.** Global and atomic agreement across Easy, Medium, and Hard difficulty tiers. **Bold**, underline, and *italics* denote first, second, and third best, respectively.

	Global				Atomic			
Method	**Easy**	**Med.**	**Hard**	**Avg.**	**Easy**	**Med.**	**Hard**	**Avg.**
CLIPScore	19.6	-3.6	-8.6	2.5	5.4	-4.8	-0.2	0.1
BLIPv2Score	21.1	3.1	-0.8	7.8	3.3	2.1	4.5	3.3
AlignScore	24.5	3.2	6.2	11.3	5.8	-0.5	0.5	1.9
PickScore	33.5	29.1	34.7	32.4	11.1	-3.2	-2.0	2.0
ImageReward	38.8	21.9	16.8	25.8	–	–	–	–
HPSv2	17.5	14.6	13.5	15.2	14.8	5.8	8.0	9.5
EXPERT	50.6	33.8	45.5	43.3	–	–	–	–
FLEUR	52.2	31.5	43.7	42.5	–	–	–	–
VIEScore	51.2	31.2	42.6	41.7	–	–	–	–
VQAScore	54.6	31.7	43.0	43.1	*52.7*	32.3	41.1	*42.0*
Qwen3.5-4B (Hol.)	48.5	39.3	40.3	42.7	–	–	–	–
Qwen3.5-4B (Dec.)	50.5	37.1	44.1	43.9	41.2	31.7	39.1	37.3
Qwen3.5-9B (Hol.)	53.7	42.2	40.4	45.4	–	–	–	–
Qwen3.5-9B (Dec.)	59.0	39.5	41.2	46.6	40.6	30.8	38.2	36.5
GPT-5.6 Sol (Hol.)	72.1	54.6	60.6	62.4	–	–	–	–
GPT-5.6 Sol (Dec.)	**76.9**	**60.7**	**62.5**	**66.7**	**70.1**	**55.6**	**68.3**	**64.7**
VisExMEM-4B	56.8	43.1	50.6	50.2	46.6	34.1	*43.2*	41.3
VisExMEM-9B	*63.3*	*45.2*	*55.6*	*54.7*	56.2	*33.0*	45.5	44.9

Table 3: **Training-Stage Ablation.** Performance of VISEXMEM across successive training stages at 4B and 9B scales. Blue rows show gains relative to the indicated reference stage; **bold** denotes the best result within each model scale.

Model	Variant	WG	EqB	TIFA	Pick	GenAI
4B	Base (Hol.)	28.2	44.0	64.3	72.5	62.2
	SFT	50.8	48.9	56.8	70.0	54.1
	vs. Base	(+22.6)	(+4.9)	(-7.5)	(-2.5)	(-8.1)
	+ CF	61.6	56.2	64.9	77.2	64.9
	vs. SFT	(+10.8)	(+7.3)	(+8.1)	(+7.2)	(+10.8)
	+ Verif.-GRPO	62.5	57.9	66.4	73.9	64.3
	+ Causal-GRPO	64.1	58.6	67.6	72.8	65.2
	+ **GDPO**	**65.6**	**59.1**	**70.2**	**80.2**	**66.3**
	vs. CF	(+4.0)	(+2.9)	(+5.3)	(+3.0)	(+1.4)
9B	Base (Hol.)	40.4	35.7	50.1	74.2	50.3
	SFT	61.9	58.9	59.8	61.8	57.2
	vs. Base	(+21.5)	(+23.2)	(+9.7)	(-12.4)	(+6.9)
	+ CF	66.3	64.9	68.9	77.5	67.2
	vs. SFT	(+4.4)	(+6.0)	(+9.1)	(+15.7)	(+10.0)
	+ Verif.-GRPO	67.4	67.2	71.5	78.0	67.9
	+ Causal-GRPO	68.5	66.7	71.0	79.1	67.1
	+ **GDPO**	**68.9**	**70.1**	**73.1**	**85.6**	**69.4**
	vs. CF	(+2.6)	(+5.2)	(+4.2)	(+8.1)	(+2.2)

Structured Evaluation Helps Across Backbones. Replacing holistic judgment with a VISEXMEM-style procedure that decomposes, verifies, and aggregates improves the average performance of Qwen3.5-4B, Qwen3.5-9B, and GPT-5.6 Sol by +5.0, +5.1, and +3.6 points, respectively, with GPT-5.6 Sol increasing from 74.0 (Hol.) to 77.6 (Dec.). These controlled comparisons show that structured inference improves average performance across all three VLM backbones, although gains vary by benchmark. Together with dedicated VISEXMEM training, the 4B and 9B models achieve the strongest open-source results in our comparison.

T2IScoreScore Performance. Figure 5 shows that both VISEXMEM variants remain in the high-ordering, high-separation region across *All, Synth, Nat, and Real*. VISEXMEM-4B performs strongly on All, Synth, and Nat, while VISEXMEM-9B shows its clearest advantage on Real, where it achieves the highest ordering score while maintaining strong separation. This indicates strong performance on naturally occurring T2I errors as well as the constructed TS2 settings.

Fine-Grained Human Alignment on AGITA-Bench. Table 2 reports agreement with human judgments at both the global and atomic levels. VISEXMEM-9B achieves the strongest open-source global average at 54.7 and improves over the corresponding untuned Qwen3.5 controls across both evaluation levels. Among methods that directly score the fixed benchmark claims, it reaches an atomic average of 44.9, compared with 64.7 for GPT-5.6 Sol (Dec.). The higher correlations on Hard than Medium should not be read as Hard being easier. In T2I, difficulty reflects semantic complexity, and Hard prompts contain more claims (15.72 vs. 8.58) and lower human alignment (2.99 vs. 3.49), which changes the within-tier score distribution. This pattern is not observed in I2T, where Medium and Hard have comparable claim counts.

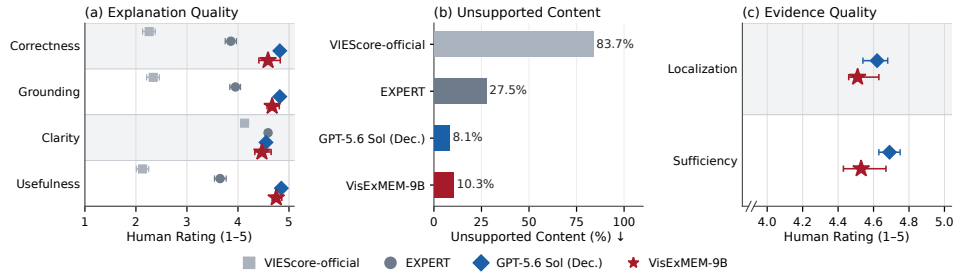

Figure 6: **Human evaluation of explanation quality.** (a) Correctness, visual grounding, clarity, and diagnostic usefulness. (b) Unsupported-content rate. (c) Localization accuracy and evidence sufficiency. Error bars show 95% confidence intervals.

Explanation Quality Analysis. Figure 6 shows that VISEXMEM-9B is close to GPT-5.6 Sol on explanation quality, reaching 4.75 diagnostic usefulness with 10.3% unsupported content, compared with 4.85 and 8.1% for GPT-5.6 Sol. It exceeds EXPERT and VIEScore in usefulness (3.65 and 2.13) while producing much less unsupported content (27.5% and 83.7%). Its localized evidence also receives high ratings for localization accuracy and sufficiency, showing that the highlighted regions are relevant and provide useful evidence for judging the associated claims.

Training-Stage Ablation. Table 3 shows that SFT improves the five-benchmark average over Base by +1.9/+9.8 points for 4B/9B, although the gains vary across benchmarks. Counterfactual refinement then gives a larger and more consistent gain over SFT of +8.8/+9.0 points. From the counterfactual checkpoint, Verification-GRPO and Causal-GRPO provide smaller, less consistent gains, while joint GDPO performs best on all five benchmarks at both scales and adds a further +3.3/+4.5 points on average. This supports the progression from structured supervision to counterfactual refinement and joint semantic and evidence-aware optimization.

Efficiency, Cost, and Reproducibility. Although GPT-5.6 Sol benefits from the decomposed evaluation procedure, evaluating the eight benchmarks and AGITA-Bench cost approximately $800 (∼$150 holistic and ∼$650 decomposed). With vLLM-based (Kwon et al., 2023) inference, VISEXMEM-9B requires 8.9 s per image-text pair including explanation generation, compared with about 30 s for GPT-5.6 Sol. At 9B parameters, VISEXMEM-9B provides a compact, locally deployable alternative to proprietary frontier judges. Its fixed checkpoint, deterministic aggregation, and explicit atomic decisions and visual evidence support interpretability and reproducibility without reliance on a proprietary API.

7 CONCLUSION, LIMITATIONS, AND FUTURE WORK

We introduced VISEXMEM, an evidence-grounded, reference-free image–text evaluator that derives a global score from localized atomic verification, together with AGITA-BENCH for evaluating global and fine-grained human agreement. Across established benchmarks, controlled VLM-as-a-Judge experiments, and human evaluation, structured decomposed prompting improves average performance across the tested backbones, while task-specific training further strengthens sensitivity to semantic errors and produces grounded explanations approaching those of proprietary judges. Overall, our results show that compact, open evaluators can combine strong alignment judgment with interpretable evidence and reproducible local inference.

VISEXMEM remains limited by errors in fine-grained verification and evidence localization, while its multi-stage inference is still more expensive than lightweight embedding-based metrics. Our prompting experiments also do not exhaust alternatives such as self- or other-feedback, and AGITA-Bench focuses on compositional semantic alignment rather than perceptual, aesthetic, or domain-specific quality. Future work should improve binding-sensitive verification and grounding, reduce inference cost, explore stronger verification strategies, and extend atomic evaluation to broader domains and modalities. More broadly, these directions could support increasingly capable, open, and auditable evaluators that reduce reliance on proprietary VLM judges.

AI USE STATEMENT

We used large language models for literature retrieval and discovery, improving the clarity and presentation of the manuscript, and providing feedback during research development. We also used Claude Code for implementation, debugging, and routine engineering tasks. Model-based components used in the research, including the offline teacher for training-data construction and model-assisted AGITA-Bench decomposition, are described in the paper. All research questions, methodological choices, experimental design, training objectives, evaluation protocols, analyses, and conclusions were determined by the authors. AI-generated suggestions, code, and text were reviewed, tested, and revised before use. We take full responsibility for the work presented in this paper.

ETHICS STATEMENT

VISEXMEM is designed to evaluate semantic alignment between images and text. Its scope does not extend to image quality, social acceptability, or the correctness of culturally or demographically sensitive representations. Like other learned multimodal evaluators, its behavior can reflect biases and representation gaps in the backbone models, grounding model, teacher models, and training data. This is especially relevant for content involving people, demographic attributes, or culturally specific concepts. Our out-of-distribution experiments on CulturalFrames and ProtoBias therefore characterize metric behavior under distribution shifts, with conclusions limited to the evaluated benchmarks. Localized visual evidence and natural-language explanations make predictions easier to inspect, but they do not guarantee correctness. Errors can arise during decomposition, grounding, or claim verification, so consequential uses should include independent human review and other sources of evidence.

Human annotation is used in AGITA-Bench and in the explanation and evidence evaluation. These tasks focus on image-text alignment, decomposition quality, explanation quality, and visual evidence. In the explanation study, system identities are hidden and outputs are shown under randomized labels to reduce presentation bias. Annotator demographics are reported only in aggregate. Our experiments rely mainly on existing public research datasets, whose collection procedures, licensing conditions, geographic and cultural coverage, and representation of people and concepts can affect the resulting evaluation. We therefore interpret the reported results within the scope of the datasets and benchmarks used in this study.

REFERENCES

Uri Berger, Gabriel Stanovsky, Omri Abend, and Lea Frermann. Surveying the landscape of image captioning evaluation: A comprehensive taxonomy, trends, and metrics analysis. *Transactions of the Association for Computational Linguistics*, 13:1597–1644, 2025. doi: 10.1162/tacl.a.52. URL https://aclanthology.org/2025.tacl-1.73/.

Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 381–389. IEEE, 2018. doi: 10.1109/WACV.2018.00048.

Wenzhi Chen, Bo Hu, Leida Li, Lihuo He, Wen Lu, and Xinbo Gao. Hyperalign: Hyperbolic entailment cones for adaptive text-to-image alignment assessment, 2026a. URL https://arxiv.org/abs/2601.04614.

Zhaorun Chen, Zichen Wen, Yichao Du, Yiyang Zhou, Chenhang Cui, Siwei Han, Jen Weng, Chaoqi Wang, Zhengwei Tong, Leria Huang, Canyu Chen, Haoqin Tu, Qinghao Ye, Zhihong Zhu, Yuqing Zhang, Jiawei Zhou, Zhuokai Zhao, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Mj-bench: Is your multimodal reward model really a good judge for text-to-image generation? In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 69174–69228. Curran Associates, Inc., 2025. doi: 10.52202/085713-2081. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/59d2eaa5842fa641ff9b8e4c7ff0f6ee-Abstract-Datasets_and_Benchmarks_Track.html.

Zheng Chen, Xun Zhang, Wenbo Li, Renjing Pei, Fenglong Song, Xiongkuo Min, Xiaohong Liu, Xin Yuan, Yong Guo, and Yulun Zhang. Grounding-iqa: Grounding multimodal language model for image quality assessment. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 39839–39851, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/42f23b6ed414b846a46c8db43509cb92-Paper-Conference.pdf.

Jaemin Cho, Yushi Hu, Jason Baldridge, Roopal Garg, Peter Anderson, Ranjay Krishna, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 15625–15645, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/446905abfec9db8856f7f7465bbd5552-Paper-Conference.pdf.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 91–104, June 2025.

Shuhao Han, Haotian Fan, Jiachen Fu, Liang Li, Tao Li, Junhui Cui, Yunqiu Wang, Yang Tai, Jingwei Sun, Chun-Le Guo, and Chongyi Li. Evalmuse-40k: A fine-grained benchmark with comprehensive human annotations for text-to-image generation model alignment evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(6):4583–4591, 2026. doi: 10.1609/aaai.v40i6.42458. URL https://ojs.aaai.org/index.php/AAAI/article/view/42458.

Sebastian Hartwig, Dominik Engel, Leon Sick, Hannah Kniesel, Tristan Payer, Poonam Poonam, Michael Glöckler, Alex Bäuerle, and Timo Ropinski. A survey on quality metrics for text-to-image generation. *IEEE Transactions on Visualization and Computer Graphics*, 31(10):9464–9483, 2025. doi: 10.1109/TVCG.2025.3585077.

Kevin Hayes, Micah Goldblum, Vikash Sehwag, Gowthami Somepalli, Ashwinee Panda, and Tom Goldstein. Finegrain: Evaluating failure modes of text-to-image models with vision language model judges. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 132119–132143. Curran Associates, Inc., 2025. doi: 10.52202/085713-3977. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/accaab47784b91d9d381868eb638be0a-Abstract-Datasets_and_Benchmarks_Track.html.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.595. URL https://aclanthology.org/2021.emnlp-main.595/.

Seyed Mohammad Hadi Hosseini, Amir Mohammad Izadi, Ali Abdollahi, Armin Saghafian, and Mahdieh Soleymani Baghshah. T2i-fineeval: Fine-grained compositional metric for text-to-image evaluation, 2025. URL https://arxiv.org/abs/2503.11481.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20406–20417, October 2023. 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.

Hang Hua, Ziyun Zeng, Yizhi Song, Yunlong Tang, Liu He, Daniel G. Aliaga, Wei Xiong, and Jiebo Luo. Mmig-bench: Towards comprehensive and explainable evaluation of multi-modal image generation models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 159706–159743. Curran Associates, Inc., 2025. doi: 10.52202/085713-4818. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/d3f8714b9f9727b43d64019f2f4043d2-Abstract-Datasets_and_Benchmarks_Track.html.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3563–3579, 2025. doi: 10.1109/TPAMI.2025.3531907. URL https://doi.org/10.1109/TPAMI.2025.3531907.

Kung-Hsiang Huang, Mingyang Zhou, Hou Pong Chan, Yi Fung, Zhenhailong Wang, Lingyu Zhang, Shih-Fu Chang, and Heng Ji. Do LVLMs understand charts? analyzing and correcting factual errors in chart captioning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 730–749, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.41. URL https://aclanthology.org/2024.findings-acl.41/.

Hyunjong Kim, Sangyeop Kim, Jongheon Jeong, Yeongjae Cho, and Sungzoon Cho. EXPERT: An explainable image captioning evaluation metric with structured explanations. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 26642–26657, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1367. URL https://aclanthology.org/2025.findings-acl.1367/.

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 A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 36652–36663. Curran Associates, Inc., 2023. doi: 10.52202/075280-1594. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/73aacd8b3b05b4b503d58310b523553c-Paper-Conference.pdf.

Shuitsu Koyama, Kazuki Matsuda, Yuiga Wada, Shinnosuke Hirano, Daichi Yashima, and Komei Sugiura. Rigel: Self-distilled score adaptation for image and video captioning evaluation, 2026. URL https://arxiv.org/abs/2606.29997.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. doi: 10.1007/s11263-016-0981-7. URL https://doi.org/10.1007/s11263-016-0981-7.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. VIEScore: Towards explainable metrics for conditional image synthesis evaluation. In Lun-Wei Ku, Andre Martins, and

Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12268–12290, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.663. URL https://aclanthology.org/2024.acl-long.663/.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, Mar 2020. ISSN 1573-1405. doi: 10.1007/s11263-020-01316-z. URL https://doi.org/10.1007/s11263-020-01316-z.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, pp. 611–626, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702297. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

Yebin Lee, Imseong Park, and Myungjoo Kang. FLEUR: An explainable reference-free evaluation metric for image captioning using a large multimodal model. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3732–3746, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.205. URL https://aclanthology.org/2024.acl-long.205/.

Christoph Leiter, Yuki M. Asano, Margret Keuper, and Steffen Eger. CROC: Evaluating and training T2I metrics with pseudo- and human-labeled contrastive robustness checks. *Transactions of the Association for Computational Linguistics*, 14:1119–1139, 2026. doi: 10.1162/tacl.a.699. URL https://aclanthology.org/2026.tacl-1.50/.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Evaluating and improving compositional text-to-visual generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 5290–5301, June 2024a. URL https://openaccess.thecvf.com/content/CVPR2024W/EvGenFM/html/Li_Evaluating_and_Improving_Compositional_Text-to-Visual_Generation_CVPRW_2024_paper.html.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation, 2024b. URL https://arxiv.org/abs/2402.17245.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Kun Li, George Vosselman, and Michael Ying Yang. Multimodal rationales for explainable visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 191–201, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025W/MULA2025/html/Li_Multimodal_Rationales_for_Explainable_Visual_Question_Answering_CVPRW_2025_paper.html.

Zhikai Li, Jiatong Li, Xuewen Liu, Wangbo Zhao, Pan Du, Kaicheng Zhou, Qingyi Gu, Yang You, Zhen Dong, and Kurt Keutzer. K-sort eval: Efficient preference evaluation for visual generation via corrected vlm-as-a-judge. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 34622–34637, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/3a81bc33b77662ed235a3328c00aebce-Paper-Conference.pdf.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://doi.org/10.1007/978-3-319-10602-1_48.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation, 2024. URL https://arxiv.org/abs/2404.01291.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. doi: 10.1162/tacl_a_00566. URL https://aclanthology.org/2023.tacl-1.37/.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Yejin Choi, Jan Kautz, and Pavlo Molchanov. Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/63333.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Computer Vision – ECCV 2024*, Lecture Notes in Computer Science, pp. 38–55. Springer, 2024. doi: 10.1007/978-3-031-72970-6_3. URL https://doi.org/10.1007/978-3-031-72970-6_3.

Mao-Lin Luo, Zhe-Xu Wang, Zi-Hao Zhou, Bo Ye, Jian Zhao, Min-Ling Zhang, and Tong Wei. Rl forgets! towards continual policy optimization, 2026. URL https://arxiv.org/abs/2607.04364.

Samin Mahdizadeh Sani, Max Ku, Nima Jamali, Matina Sani, Paria Khoshtab, Wei-Chieh Sun, Parnian Fazel, Zhi Rui Tam, Thomas Chong, Edisy Kin Wai Chan, Donald Tsang, Chiao-Wei Hsu, Ting Lam, Ho Yin Sam Ng, Chiafeng Chu, Chak-Wing Mak, Keming Wu, Hiu Tung Wong, Yik Chun Ho, Chi Ruan, Zhuofeng Li, I-Sheng Fang, Shih-Ying Yeh, Ho Kei Cheng, Ping Nie, and Wenhu Chen. Imagenworld: Stress-testing image generation models with explainable human evaluation on open-ended real-world tasks. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 41886–41916, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/458fa8ee331566383d8e74bdb647f829-Paper-Conference.pdf.

Sina Malakouti, Boqing Gong, and Adriana Kovashka. Culture in action: Evaluating text-to-image models through social activities. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 136235–136262, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/dca63f2650fe9e88956c1b68440b8ee9-Paper-Conference.pdf.

Shravan Nayak, Mehar Bhatia, Xiaofeng Zhang, Verena Rieser, Lisa Anne Hendricks, Sjoerd van Steenkiste, Yash Goyal, Karolina Stanczak, and Aishwarya Agrawal. CulturalFrames: Assessing cultural expectation alignment in text-to-image models and evaluation metrics. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 20918–20953, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1141. URL https://aclanthology.org/2025.findings-emnlp.1141/.

Yuansheng Ni, Songcheng Cai, Xiangchao Chen, Jiarong Liang, Zhiheng LYU, Jiaqi Deng, Kai Zou, PING NIE, Fei Yuan, Xiang Yue, and Wenhu Chen. Viscoder2: Building multi-language visualization coding agents. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust

(eds.), *International Conference on Learning Representations*, volume 2026, pp. 104353–104422, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/aac47a65339d7353562f6365526ee81b-Paper-Conference.pdf.

Masanari Oi, Masahiro Kaneko, Naoaki Okazaki, and Nakamasa Inoue. Multi-modal, multi-task, multi-criteria automatic evaluation with vision language models. In Stelios Piperidis, Núria Bel, Henk van den Heuvel, Nancy Ide, Simon Krek, and Antonio Toral (eds.), *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pp. 9396–9408, Palma de Mallorca, Spain, May 2026. ELRA Language Resource Association. doi: 10.63317/575aknbvd9hs. URL https://aclanthology.org/2026.lrec-1.738/.

Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, Su Wang, and Jason Baldridge. Docci: Descriptions of connected and contrasting images, 2024. URL https://arxiv.org/abs/2404.19753.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3170–3180, October 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Paiss_Teaching_CLIP_to_Count_to_Ten_ICCV_2023_paper.html.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2641–2649, December 2015. doi: 10.1109/ICCV.2015.303. URL https://openaccess.thecvf.com/content_iccv_2015/html/Plummer_Flickr30k_Entities_Collecting_ICCV_2015_paper.html.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, volume 12350 of *Lecture Notes in Computer Science*, pp. 647–664, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58558-7. doi: 10.1007/978-3-030-58558-7_38. URL https://doi.org/10.1007/978-3-030-58558-7_38.

Shu Pu, Yaochen Wang, Dongping Chen, Yuhang Chen, Guohao Wang, Qi Qin, Zhongyi Zhang, Zhiyuan Zhang, Zetong Zhou, Shuang Gong, Yi Gui, Yao Wan, and Philip S. Yu. Judge anything: Mllm as a judge across any modality. In *KDD '25: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 5742–5753, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714542. doi: 10.1145/3711896.3737409. URL https://doi.org/10.1145/3711896.3737409.

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

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13009–13018, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Rasheed_GLaMM_Pixel_Grounding_Large_Multimodal_Model_CVPR_2024_paper.html.

Candace Ross, Melissa Hall, Adriana Romero Soriano, and Adina Williams. What makes a good metric? evaluating automatic metrics for text-to-image consistency. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=LFfktMPAci.

Subhadeep Roy, Gagan Bhatia, and Steffen Eger. Prototypicality bias reveals blindspots in multimodal evaluation metrics, 2026. URL https://arxiv.org/abs/2601.04946.

Sara Sarto, Marcella Cornia, and Rita Cucchiara. Image captioning evaluation in the age of multimodal llms: Challenges and future perspectives. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 10632–10640. International Joint Conferences on Artificial Intelligence Organization, 8 2025. doi: 10.24963/ijcai.2025/1180. URL https://doi.org/10.24963/ijcai.2025/1180. Survey Track.

Michael Saxon, Fatima Jahara, Mahsa Khoshnoodi, Yujie Lu, Aditya Sharma, and William Yang Wang. Who evaluates the evaluations? objectively scoring text-to-image prompt coherence metrics with t2iscorescore (ts2). In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 85630–85657. Curran Associates, Inc., 2024. doi: 10.52202/079017-2719. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/9b9cfd5428153ccfbd4ba34b7e007305-Paper-Conference.pdf.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8430–8439, October 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Shao_Objects365_A_Large-Scale_High-Quality_Dataset_for_Object_Detection_ICCV_2019_paper.html.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

FuLin Shi, Wenyi Xiao, Leilei Gan, Liang Ding, and Binchen. REVEALER: Reinforcement-guided visual reasoning for element-level text-image alignment evaluation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 47630–47649, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.2200. URL https://aclanthology.org/2026.acl-long.2200/.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5238–5248, June 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Thrush_Winoground_Probing_Vision_and_Language_Models_for_Visio-Linguistic_Compositionality_CVPR_2022_paper.html.

Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Equivariant similarity for vision-language foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11998–12008, October 2023a. URL https://openaccess.thecvf.com/content/ICCV2023/html/Wang_Equivariant_Similarity_for_Vision-Language_Foundation_Models_ICCV_2023_paper.html.

Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 893–911, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.51. URL https://aclanthology.org/2023.acl-long.51/.

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.

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 A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 15903–15935. Curran Associates, Inc., 2023. doi: 10.52202/075280-0700. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/33646ef0ed554145eab65f6250fab0c9-Paper-Conference.pdf.

Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. What you see is what you read? improving text-image alignment evaluation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 1601–1619. Curran Associates, Inc., 2023. doi: 10.52202/075280-0079. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/056e8e9c8ca9929cb6cf198952bf1dbb-Paper-Conference.pdf.

Gyeongtae Yoo, Sanghyeok Park, Soohyuk Jang, Ik hwan Kim, and Sungroh Yoon. Rollout-level advantage-prioritized experience replay for grpo, 2026. URL https://arxiv.org/abs/2606.04560.

Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8018–8027, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Zhang_Learning_Multi-Dimensional_Human_Preference_for_Text-to-Image_Generation_CVPR_2024_paper.html.

Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. Gpt-4v(ision) as a generalist evaluator for vision-language tasks, 2023. URL https://arxiv.org/abs/2311.01361.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 633–641, July 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Zhou_Scene_Parsing_Through_CVPR_2017_paper.html.

A Appendix

A.1 Training Data Construction

Data sources. Table 4 summarizes the public datasets used to construct the training corpus and the supervision extracted from each source. We combine dense region annotations, grounded captions, spatial-relation labels, count annotations, natural long-form descriptions, human judgments, and contrastive image pairs to obtain complementary coverage of the semantic skills required by VisExMem.

Atomic claim construction. We represent each caption as a set of typed, visually verifiable claims drawn from six categories: object, attribute, action, relation, count, and negation. When the source dataset provides explicit structure–e.g., Visual Genome objects and attributes, Flickr30k phrase–region correspondences, VSR spatial labels, or PixMo-Count annotations–we convert this supervision programmatically. For sources dominated by free-form descriptions, including DOCCI, CROC, and Localized Narratives, we use `Qwen3.6-27B`[1] as an offline decomposition teacher. The teacher is not used to overwrite reliable source annotations.

[1] GPT-5.6 Sol, Qwen3.6-27B

Atomic verification. Each claim receives a supported or contradicted decision. Labels are derived preferentially from native dataset annotations or human judgments. When a reliable decision cannot be obtained directly from the source, the offline teacher may provide a missing decision or structural link; low-confidence or unresolved outputs are discarded. Thus, teacher supervision supplements rather than replaces source-grounded labels.

Counterfactual supervision. We construct counterfactual examples through two complementary mechanisms. First, CROC provides image-side contrastive pairs in which the same textual description is evaluated against independently generated images with known match labels. Second, for verified positive claims on natural images, we programmatically create text-side counterfactuals by modifying a single semantic constraint—object, attribute, action, relation, count, or negation—while preserving the image. This produces minimal pairs in which the intended semantic change is explicitly known.

Visual evidence. Evidence is taken directly from native annotations whenever available. For claims without native bounding boxes, we use GroundingDINO to localize the already-labeled semantic content. GroundingDINO therefore serves only as an evidence-localization mechanism and never provides the Yes/No correctness label. A failed detection is treated as a localization failure rather than evidence that the claim is absent. Evidence is represented according to claim type: object and attribute claims use the relevant regions, relations and actions may involve multiple participating regions, and counting claims retain the set of candidate instances. Pure absence claims are not assigned fabricated evidence regions.

Causal-evidence construction. To distinguish genuine visual evidence from arbitrary image regions, each eligible evidence region E_i is paired with a random control region R_i from the same image. These selected-versus-random pairs provide the controlled interventions used by the Stage 3 causal-evidence objective. Exact random-region matching and image-view construction are described in Appendix A.2.5.

Explanation supervision. Explanation supervision is constructed only after the atomic claims, decisions, and evidence have been fixed. Explanation targets are generated from these structured records rather than used to determine them. In particular, the explanation teachers do not infer new bounding boxes or redefine correctness labels. This separation ensures that explanation learning cannot modify the supervision from which the numerical metric score is derived.

Processed training examples. Figure 7 shows representative trainer-facing records after preprocessing. Stage 1 uses separate SFT views for decomposition, verification, evidence selection, and explanation. GroundingDINO boxes are supplied as candidate evidence regions; the evidence-selection task predicts candidate IDs rather than box coordinates. Stage 2 uses paired supported and counterfactual claims and optimizes their decision margins without a generated target sequence. Stage 3 combines semantic examples, where groups of sampled completions are scored by Yes/No correctness, with causal-evidence examples that compare selected evidence against a matched random region using multiple image views. Stage 3 also replays earlier SFT and counterfactual examples. The examples below are taken from the processed training files; unused bookkeeping fields are omitted and long coordinates are rounded only for readability.

Data isolation. All final evaluation benchmarks are excluded from the training corpus and model-selection data. Training and development groups are separated before optimization to prevent the same underlying example from appearing across splits. Only the training split of EvalMuse is used for training.

A.2 Training and Implementation Details

A.2.1 Models and Parameter-Efficient Training

We train Qwen3.5-4B and Qwen3.5-9B using LoRA (Hu et al., 2022) with rank 16, scaling factor $\alpha = 32$, dropout 0.05, and no bias. LoRA is applied to the q, k, v, o, gate, up, and down projections. All training uses bfloat16 precision and disables the Qwen3.5 thinking mode. The main

Stage 1: Structured SFT

```
Source: coco_train2017
Sample: coco_3333

DECOMPOSITION VIEW

Caption:
"A fire truck near a yellow fire hydrant."

Target:
{"parts":[
 {"text":"fire truck",
  "type":"object"},
 {"text":"yellow fire hydrant",
  "type":"attribute"},
 {"text":"fire truck near a yellow fire
 hydrant",
  "type":"relation"}
]}
```

```
EVIDENCE-SELECTION VIEW

Claim (attribute):
yellow fire hydrant

Candidate E1:
box=[93,279,224,482]

Target:
{"selected_evidence_ids":["E1"]}

Training signal:
supervised structured target
```

Stage 2: Counterfactual Refinement

```
Source: coco_train2017
Claim type: attribute

SUPPORTED SIDE

Caption:
"A fire truck near a yellow fire hydrant."

Claim:
yellow fire hydrant

Decision: YES
```

```
COUNTERFACTUAL SIDE

Caption:
"A fire truck near a green fire hydrant."

Claim:
green fire hydrant

Decision: NO

Change: yellow -> green
Image pixels: identical for this example

Training signal:
Yes/No decision-margin loss
Generated target text: none
```

Stage 3: Evidence-Aware GDPO

```
SEMANTIC EXAMPLE

Source: coco_train2017
Sample: cf_coco_9065_count

Caption:
"A dog and three people walking on sand."

Claim (count): three
Decision: NO

Rollouts per prompt: 8

Reward:
 +1.00  correct decision
 -1.00  incorrect decision
 -1.25  invalid JSON
```

```
CAUSAL-EVIDENCE EXAMPLE

Source: coco_train2017
Sample: coco_3333

Claim (attribute):
yellow fire hydrant

Decision: YES

Selected region:
[92.85,278.53,224.34,482.07]

Matched random region:
[141.57,62.35,270.76,272.01]

Image views:
full
mask_selected
mask_random
evidence_only
random_only
```

Figure 7: **Representative processed training examples.** Real trainer-facing examples from the three VisExMEM training stages. Unused metadata are omitted and coordinates are rounded for readability.

Source	Supervision used	Primary skills
MS COCO	Captions and object boxes	Object, count
Visual Genome	Region descriptions, object boxes, attributes	Object, attribute, relation
VSR	Image–caption pairs with spatial-relation labels	Spatial relation
Flickr30k Entities	Phrase–entity–region correspondences	Object, grounding
Open Images + Localized Narratives	Grounded free-form narratives	Object, attribute, action, relation
DOCCI	Detailed connected and contrasting descriptions	Attribute, relation, negation
PixMo-Count	Exact counts with point-level evidence	Count
CROCsyn	Contrastive image pairs with match labels	Counterfactual verification
EvalMuse-40K	Element-level human judgments from the training split	Atomic verification

Table 4: Public data sources used to construct the VISEXMEM training corpus.

Task	Target format
Decomposition	`{"parts": [...] }`; each part contains `part_id`, `text`, `type`, `parent_part_id`, `head_part_id`, and `target_part_id`.
Verification	`{"decision": "yes"}` or `{"decision": "no"}`.
Evidence selection	`{"selected_evidence_ids": ["E1", ...]}`.
Explanation	`{"explanation": "..."}`.

Table 5: Task-specific target formats used during Stage 1 SFT. Candidate bounding boxes are inputs to evidence selection and are not prediction targets.

training hyperparameters are summarized in Table 6; construction of the underlying training corpus is described separately in Appendix A.1.

A.2.2 VERIFICATION PROMPT AND DECISION SCORING

Counterfactual refinement and Stage 3 use the same verification prompt template to avoid prompt drift between supervised and policy optimization. Given an image, caption, claim type, and atomic claim, the model is instructed to return exactly one JSON decision, `{"decision":"yes"}` or `{"decision":"no"}`. Outputs that do not produce either valid decision are treated as invalid during Stage 3 policy optimization.

During counterfactual refinement, no free-form generation is performed. For each supported/counterfactual pair, the same image is evaluated with the corresponding claim, and the verification margin in Equation 2 is obtained directly from the `yes` and `no` decision logits after their shared output prefix. In Stage 3 semantic optimization, the same prompt is instead used to sample $G = 8$ decision completions. For causal-evidence training, the textual prompt is held fixed while only the image view changes. Thus, differences between causal views arise from the image intervention rather than from changes to the textual instruction.

A.2.3 MULTI-TASK SFT

Stage 1 uses a four-part curriculum over decomposition, verification, evidence selection, and joint supervision. The task-specific views contain 13.5K decomposition, 15K verification, and 7K evidence-selection examples. The final joint stage contains 40K examples, including 4.5K explanation examples, and samples decomposition, verification, evidence-selection, and explanation tasks using an 8:9:4:3 schedule. Each curriculum stage is traversed once.

SFT target formats. Each SFT example has a single task-specific JSON target y, and loss is computed only over the target tokens using the next-token objective in Equation 1. Table 5 summarizes the exact target formats. Candidate bounding boxes are provided only as inputs for evidence selection and are never prediction targets.

Optimization. We use AdamW with cosine learning-rate decay, 3% warmup, and gradient clipping at 1.0. The learning rate is 10^{-5} for the first three task-specific curriculum stages and 7×10^{-6} for final joint training.

A.2.4 COUNTERFACTUAL REFINEMENT

Stage 2 starts from the promoted Stage 1 checkpoint. The source pool contains 20K supported/counterfactual pairs. Training interleaves counterfactual updates, verification replay, and decomposition replay in an 8:1:1 update ratio.

For pair i, let $m_i^+ = m_\theta(I_i, a_i^+)$ and $m_i^- = m_\theta(I_i, a_i^-)$ denote the verification margins from Equation 2 for the supported and counterfactual claims, respectively. Let $m_{i,\mathrm{ref}}^+$ denote the positive margin produced by the frozen pre-Stage-2 checkpoint. We define the retained positive target as

$$\bar{m}_i^+ = \max\Big(\tau_+, m_{i,\mathrm{ref}}^+ - \delta_{\mathrm{keep}}\Big), \tag{5}$$

and optimize

$$\mathcal{L}_{\mathrm{CF}} = \mathbb{E}_i\Big[\lambda_{\mathrm{pair}}\left[\gamma_{\mathrm{cf}} - (m_i^+ - m_i^-)\right]_+ + \lambda_+\left[\bar{m}_i^+ - m_i^+\right]_+ + \lambda_-\left[m_i^- + \tau_-\right]_+\Big], \tag{6}$$

where $[z]_+ = \max(0, z)$. The three terms respectively enforce separation between the supported and counterfactual claims, preserve sufficient support for the positive claim, and push the counterfactual below the negative target.

We use $\gamma_{\mathrm{cf}} = 3.0$, $\tau_+ = 3.0$, $\tau_- = 2.0$, $\delta_{\mathrm{keep}} = 1.0$, $\lambda_{\mathrm{pair}} = 1.0$, and $\lambda_+ = \lambda_- = 0.5$. Thus, the positive target is $\max(3.0, m_{i,\mathrm{ref}}^+ - 1.0)$ and the negative target is -2.0.

Verification replay uses its original binary verification objective together with the retention constraint, whereas decomposition replay retains the Stage 1 token-level sequence objective. We train for one counterfactual epoch (450 optimizer steps) with AdamW, learning rate 5×10^{-7}, 4% warmup, cosine decay to a 10% learning-rate floor, and gradient clipping at 1.0.

A.2.5 EVIDENCE-AWARE GDPO

Stage 3 starts from the promoted counterfactual checkpoint and uses 20K semantic-verification examples together with 20K supported causal-evidence examples. Our final Stage 3 procedure is an evidence-aware GDPO variant: semantic correctness supplies the group-relative rollout signal, while causal evidence is optimized directly as an auxiliary objective. In particular, the causal signal is not treated as a second independently group-normalized rollout reward; this distinction is important because the image-intervention signal is shared across rollouts of the same input and would otherwise be attenuated by within-group centering.

Semantic policy optimization. For each semantic prompt, we sample $G = 8$ completions with temperature 0.7, top-$p = 1.0$, and a maximum of 16 generated tokens. Correct, incorrect, and invalid decisions receive rewards $+1$, -1, and -1.25, respectively.

The group-relative quantities in Equation 3 are implemented with

$$A_{i,g} = \frac{r_{i,g} - \mu_i}{\sigma_i + \epsilon_A}, \qquad \epsilon_A = 10^{-4}, \tag{7}$$

where μ_i and σ_i are the mean and population standard deviation of the G rollout rewards for semantic example i. If all rewards in a group are identical, every centered reward is zero and the group therefore provides no relative policy signal.

For the clipped semantic objective in Equation 4, we use $\epsilon = 0.2$. KL regularization uses $\beta = 0.02$ with the frozen Stage 2 checkpoint as the reference policy. For a generated token, the token-level KL penalty is

$$d^{\mathrm{KL}} = \exp(\kappa) - \kappa - 1, \qquad \kappa = \log \pi_{\mathrm{ref}} - \log \pi_\theta. \tag{8}$$

Replay schedule. Stage 3 interleaves semantic GDPO updates, counterfactual replay, and SFT replay in a 7:2:1 update ratio. Replay draws from the earlier training pools and introduces no additional training examples. Counterfactual replay therefore retains the Stage 2 margin objective in Equation 6, while SFT replay retains the original sequence-supervision objective in Equation 1.

Five-view causal construction. For causal example i, let E_i denote the selected evidence region and R_i its matched random control. We construct five image views: the intact image I, selected-evidence removal $I \setminus E_i$, random-control removal $I \setminus R_i$, selected-evidence-only $I \odot E_i$, and random-control-only $I \odot R_i$. Here $\setminus$ masks the specified region while retaining the rest of the image, whereas $\odot$ masks the remaining canvas and retains only the specified region. The masking implementation uses either Gaussian blur or local-mean filling.

Let $s_i(V)$ denote the `yes–no` verification margin for claim i under image view V. We define

$$\begin{aligned} \Delta_i^{\text{sel}} &= s_i(I) - s_i(I \setminus E_i), & \Delta_i^{\text{rnd}} &= s_i(I) - s_i(I \setminus R_i), \\ \Delta_i^{\text{suff}} &= s_i(I \odot E_i) - s_i(I \odot R_i), & \Delta_i^{\text{anchor}} &= s_i(I) - s_i(I \odot E_i). \end{aligned} \tag{9}$$

The first two quantities measure the effect of removing selected versus random regions, Δ_i^{suff} compares their evidence-only sufficiency, and Δ_i^{anchor} measures the drop from the intact image to the selected-evidence-only view.

Causal-evidence objective. Using the quantities in Equation 9, we optimize

$$\begin{aligned} \mathcal{L}_{\text{causal}}(u) = \lambda_{\text{ramp}}(u) \big[& w_{\text{comp}} \operatorname{softplus}\big(\gamma_{\text{comp}} - (\Delta_i^{\text{sel}} - \Delta_i^{\text{rnd}})\big) \\ & + w_{\text{drop}} \operatorname{softplus}\big(\gamma_{\text{drop}} - \Delta_i^{\text{sel}}\big) \\ & + w_{\text{suff}} \operatorname{softplus}\big(\gamma_{\text{suff}} - \Delta_i^{\text{suff}}\big) \\ & + w_{\text{suffdrop}} \operatorname{softplus}\big(\Delta_i^{\text{anchor}} - \gamma_{\max}\big) \big]. \end{aligned} \tag{10}$$

The four terms respectively encourage (i) a larger removal effect for selected than random evidence, (ii) sensitivity to removing selected evidence, (iii) greater support from selected than random evidence-only views, and (iv) a bounded degradation when verification is restricted to selected evidence alone.

We use $\gamma_{\text{comp}} = 0.25$, $\gamma_{\text{drop}} = 0.25$, $\gamma_{\text{suff}} = 0.75$, and $\gamma_{\max} = 1.0$, with

$$(w_{\text{comp}}, w_{\text{drop}}, w_{\text{suff}}, w_{\text{suffdrop}}) = (1.0, 0.5, 1.0, 0.25).$$

The outer causal coefficient in the Stage 3 objective is $\lambda_{\text{c}} = 0.3$. The internal coefficient $\lambda_{\text{ramp}}(u)$ increases linearly from 0 to 0.5 over the first 100 completed causal updates and remains at 0.5 thereafter. The maximum resulting multiplier on the weighted causal auxiliary is therefore $0.3 \times 0.5 = 0.15$.

Eligibility and random-control matching. A causal example contributes to $\mathcal{L}_{\text{causal}}$ only when the source image and all five required views can be constructed. Examples that fail this requirement are excluded from the causal auxiliary; if no eligible causal examples remain in an update, the causal term is omitted for that update.

Each selected evidence region E_i is paired with a random control region R_i from the same image. The control is matched to E_i within $\pm 15\%$ in area and $\pm 20\%$ in aspect ratio while satisfying $\text{IoU}(R_i, E) < 0.10$ for every evidence region E in that image. We allow at most 200 sampling attempts.

Joint Stage 3 objective and optimization. The final Stage 3 update combines the semantic objective from Equation 4 with the direct causal auxiliary from Equation 10:

$$\mathcal{L}_{\text{Stage3}} = -\mathcal{J}_{\text{sem}} + \lambda_{\text{c}} \mathcal{L}_{\text{causal}}. \tag{11}$$

The causal term is optimized directly rather than incorporated into the group-normalized rollout reward, preserving the image-intervention signal while semantic correctness is learned through the rollout-based policy objective.

Stage 3 uses AdamW with learning rate 10^{-6}, cosine decay, 3% warmup, zero weight decay, and gradient clipping at 1.0. The sequential Verification-GRPO and Causal-GRPO alternatives used in the training ablations are described separately in Appendix A.3; they branch from the Stage 2 checkpoint and are not additional stages of the final VISEXMEM model.

Table 6: Training hyperparameters for the 4B and 9B VISEXMEM models. For SFT, $4/2$ denotes per-device batch size 4 for the decomposition curriculum stage and 2 for the remaining stages; the two learning rates correspond to the first three task-specific curriculum stages and the final joint stage, respectively. Replay/update mixtures denote counterfactual:verification:decomposition for Stage 2 and semantic-GDPO:counterfactual:SFT for Stage 3.

	SFT		**Counterfactual**		**Evidence-aware GDPO**	
Hyperparameter	**4B**	**9B**	**4B**	**9B**	**4B**	**9B**
Training duration (hours)	4.4	4.6	2.7	3.4	40.9	43.3
GPUs	$4\times$ H100	$4\times$ H100	$4\times$ H100	$4\times$ H100	$2\times$ H100	$2\times$ H100
Per-device batch size	$4/2$	$4/2$	1	1	2	2
Gradient accumulation	8	8	8	8	2	2
Learning rate	$10^{-5}/7\times10^{-6}$	$10^{-5}/7\times10^{-6}$	5×10^{-7}	5×10^{-7}	10^{-6}	10^{-6}
Learning-rate scheduler	cosine	cosine	cosine	cosine	cosine	cosine
Warmup	3%	3%	4%	4%	3%	3%
Weight decay	0.0	0.0	0.01	0.01	0.0	0.0
Gradient clipping	1.0	1.0	1.0	1.0	1.0	1.0
Replay/update mixture	–	–	8:1:1	8:1:1	7:2:1	7:2:1
Generations per prompt	–	–	–	–	8	8

A.3 GRPO ABLATIONS

Ablation strategy. We evaluate a sequential GRPO alternative to the final VISEXMEM training strategy. The ablation branch starts from the counterfactually refined Stage 2 checkpoint:

$$\text{Counterfactual} \rightarrow \text{Verification-GRPO} \rightarrow \text{Causal-GRPO}.$$

In contrast, the final evidence-aware GDPO stage branches directly from the counterfactual checkpoint. Thus, Verification-GRPO and Causal-GRPO are controlled ablations rather than additional stages of the deployed model.

Verification-GRPO. Verification-GRPO isolates reinforcement learning for semantic correctness. For each image–claim input, we sample $G = 8$ completions and assign reward $+1$ to a correct `yes`/`no` decision, -1 to an incorrect valid decision, and -1.25 to an invalid output. Rewards are normalized within each rollout group as

$$A_{i,g} = \frac{r_{i,g} - \mu_i}{\sigma_i + \epsilon_A}, \qquad \epsilon_A = 10^{-4}, \tag{12}$$

where μ_i and σ_i are the mean and population standard deviation of the G rollout rewards. Groups with identical rewards therefore provide no relative policy signal. We then apply the same clipped policy objective and KL regularization described in Equation 4. Unlike Stage 2, which directly optimizes verification margins, Verification-GRPO learns from relative differences between sampled decisions.

Causal-GRPO. Causal-GRPO continues from Verification-GRPO and tests whether visual evidence dependence can be incorporated through the rollout reward itself. It uses the same GRPO formulation, but for a correct supported decision adds a causal bonus derived from the difference between masking selected evidence and a matched random region:

$$r_{i,g}^{\text{causal}} = r_{i,g}^{\text{sem}} + \alpha \ \text{clip}(e_i, 0, 1), \qquad \alpha = 0.3, \tag{13}$$

where e_i denotes the selected-versus-random evidence advantage. This ablation therefore tests a *reward-based* treatment of causal evidence, whereas the final model optimizes evidence dependence directly.

Replay. Both GRPO ablations interleave decomposition examples from the earlier SFT stage to preserve the structured evaluation interface during policy optimization. This differs from GRPO experience replay (Yoo et al., 2026): we do not store or reuse previous policy rollouts. Instead, earlier supervised examples are revisited with their original training objective.

The final GDPO stage uses the same retention principle more broadly, interleaving semantic policy updates, counterfactual replay, and SFT replay. The contribution of this replay strategy is evaluated separately in the training ablations.

Relation to the final GDPO objective. The sequential GRPO branch separates semantic verification and causal evidence into two successive policy-optimization stages. Our final model instead combines them in a single Stage 3 optimization. Semantic verification provides the rollout-based policy signal, while the five-view causal objective from Equation 10 is optimized directly:

$$\mathcal{L}_{\text{Stage3}} = -\mathcal{J}_{\text{sem}} + \lambda_{\text{c}}\mathcal{L}_{\text{causal}}. \tag{14}$$

This distinction is important because a causal signal shared across all rollouts of the same input is largely removed by within-group centering and therefore provides little relative GRPO signal. Direct optimization preserves the image-intervention signal while allowing semantic correctness and evidence dependence to be learned jointly.

A.4 Additional Results

A.4.1 Out-of-Distribution Generalization

We further evaluate VisExMEM on three external benchmarks that are not used in training, covering structured visualizations, culturally grounded image–text alignment, and prototypicality-sensitive semantic evaluation.

VisPlotBench (Ni et al., 2026), introduced with VisCoder2, provides a structured-visual distribution based on rendered visualization tasks. We evaluate 886 examples and report pairwise accuracy (PAcc) and the stricter Group score. We additionally include ChartVE (Huang et al., 2024), a visual-entailment model developed for factual-consistency evaluation of chart captions.

CulturalFrames (Nayak et al., 2025) evaluates alignment with explicit and implicit cultural expectations across countries and socio-cultural domains. We evaluate 3,577 images and report Spearman's ρ with human alignment judgments and within-prompt pairwise accuracy (PAcc_{wp}). We also include the culturally targeted AHEaD metrics (Malakouti et al., 2026): ALIGN measures coverage of expected cultural descriptors, while HAL measures unsupported cultural content. We report $1 - \text{HAL}$ so that higher is better throughout.

ProtoBias (Roy et al., 2026) tests whether evaluation metrics prefer semantically correct images over more prototypical alternatives containing controlled semantic errors. We evaluate its 600 human-annotated examples, with 200 each from Animals, Demography, and Objects, and report category-level accuracy and Cohen's κ. The GPT-5 (Judge) and ProtoScore results are taken directly from the original ProtoBias study (Roy et al., 2026) and are marked with † in the table; all remaining values are obtained under our evaluation setup.

Results. The OOD results show that VisExMEM-specific training provides substantial gains over the corresponding untuned Qwen3.5 judges. On VisPlotBench, VisExMEM-4B improves over both 4B controls and reaches 91.3 PAcc and 74.4 Group, while VisExMEM-9B reaches 91.3 PAcc and the highest Group score of 78.6, compared with 74.5 for the strongest 9B Qwen control. On CulturalFrames, the gain is clearest at 9B: VisExMEM-9B reaches $\rho = .311$ and 65.8 PAcc_{wp}, exceeding both corresponding Qwen3.5 controls and all conventional metrics; the strongest conventional baseline, VQAScore, obtains $\rho = .250$ and 64.4 PAcc_{wp}.

The same pattern appears on ProtoBias. VisExMEM-9B substantially improves over both 9B Qwen controls across all three categories, reaching 73.5/71.5/81.0 accuracy and .254/.205/.286 κ on Animals/Demography/Objects. VisExMEM-4B likewise improves strongly over its Qwen baselines in category accuracy. Among the conventional general-purpose metrics, the VisExMEM variants obtain the strongest ProtoBias results. ProtoScore achieves higher category accuracy, as expected for a model trained specifically for the prototypicality-bias setting, while VisExMEM-9B remains stronger in κ on Animals and Objects despite receiving no ProtoBias-specific training.

Table 7: **Out-of-distribution generalization.** Results on VisPlotBench, CulturalFrames, and the 600 human-annotated ProtoBias examples. VisPlotBench reports pairwise accuracy (PAcc) and Group score; CulturalFrames reports Spearman's ρ and within-prompt pairwise accuracy (PAcc_{wp}); ProtoBias reports category-level accuracy (Acc.) and Cohen's κ. Accuracy-based metrics are reported on a 0–100 scale; correlations and κ are reported on their native scale. Higher is better. Bold, underlined, and italicized values denote the best, second-best, and third-best results in each column, respectively.

Method	VisPlotBench		CulturalFrames		ProtoBias					
					Animal		Demography		Object	
	PAcc	Group	ρ	PAcc_{wp}	Acc.	κ	Acc.	κ	Acc.	κ
CLIPScore	82.5	52.6	.145	60.5	25.0	.001	28.0	-.005	32.0	.058
ALIGNScore	79.1	45.3	.149	57.8	51.0	.033	34.0	-.064	32.0	.011
PickScore	77.1	39.5	.206	62.2	32.0	.001	34.0	-.014	31.0	.126
HPSv2	79.7	37.4	.219	61.5	25.0	.024	38.0	-.006	38.0	.125
VQAScore	85.1	61.4	.250	64.4	51.0	.028	47.0	-.007	55.0	.065
ChartVE	69.2	31.3	–	–	–	–	–	–	–	–
GPT-5 (Judge)	–	–	–	–	56.0	.108	51.0	*.176*	54.0	*.265*
ProtoScore	–	–	–	–	**95.0**	*.182*	**91.0**	**.232**	**98.0**	.178
AHEaD-ALIGN	–	–	.114	55.6	35.5	.064	49.5	.083	55.5	-.030
AHEaD-HAL (1 − HAL)	–	–	.131	55.0	–	–	–	–	–	–
Qwen3.5-4B (Hol.)	84.4	70.2	.242	56.7	39.8	.131	45.8	.075	44.5	.161
Qwen3.5-4B (Dec.)	85.8	73.8	.296	49.4	43.5	.121	39.0	.102	60.0	.269
Qwen3.5-9B (Hol.)	81.5	72.4	*.252*	15.2	32.0	.049	26.0	.142	40.5	.102
Qwen3.5-9B (Dec.)	83.4	74.5	.210	51.4	42.9	.056	54.0	.167	55.6	.195
VisExMEM-4B	**91.3**	*74.4*	.198	60.5	76.0	.189	*67.0*	.106	*73.5*	.256
VisExMEM-9B	**91.3**	**78.6**	**.311**	**65.8**	*73.5*	**.254**	71.5	.205	81.0	**.286**

Table 8: **Replay ablation for VISEXMEM-9B.** Both paths start from the same SFT checkpoint and either retain or disable replay throughout CF and GDPO. Δ denotes Replay − No replay. Higher is better.

Variant	Replay	WG	EqB	TIFA	Pick	GenAI
Base (Hol.)	–	40.4	35.7	50.1	74.2	50.3
+ SFT	–	61.9	58.9	59.8	61.8	57.2
+ CF	**Yes**	66.3	64.9	68.9	77.5	67.2
+ CF	**No**	**67.9**	**66.2**	**71.1**	**81.1**	**67.3**
Δ	*Yes − No*	(-1.6)	(-1.3)	(-2.2)	(-3.6)	(-0.1)
+ GDPO	**Yes**	**68.9**	**70.1**	**73.1**	**85.6**	**69.4**
+ GDPO	**No**	66.1	64.8	70.7	74.5	65.1
Δ	*Yes − No*	(+2.4)	(+5.3)	(+2.4)	(+11.1)	(+4.3)

A.4.2 REPLAY ABLATION

We examine the role of replay during progressive training using VISEXMEM-9B. Both variants start from the same base model and SFT checkpoint. We then train two complete trajectories: the standard trajectory replays earlier-stage supervision during both counterfactual refinement (CF) and evidence-aware GDPO, whereas the no-replay trajectory disables replay in both stages. Table 8 reports the resulting checkpoints on the same five benchmarks used in the main training-stage ablation.

At the CF checkpoint, disabling replay is consistently beneficial. The no-replay variant improves over the replay variant on all five benchmarks, raising the average from 69.0 to 70.7. The largest difference occurs on Pick-a-Pic (81.1 vs. 77.5), while GenAI-Bench changes only marginally (67.3 vs. 67.2). Thus, replay is not required for effective counterfactual adaptation and slightly constrains the immediate CF checkpoint.

The behavior changes markedly after evidence-aware GDPO. Starting from the stronger no-replay CF checkpoint, continuing training without replay reduces the average from 70.7 to 68.2. In contrast,

Table 9: **Aggregation ablation.** Performance of VISEXMEM-9B under alternative fixed aggregation rules. Higher is better.

Aggregation	WG	EqBen	TIFA160	Pick-a-Pic	GenAI
Arithmetic mean	**68.9**	**70.1**	**73.1**	**85.6**	**69.4**
Geometric	68.4	69.6	71.6	84.4	68.7
Harmonic	67.3	60.3	69.6	81.5	64.4
Minimum	62.9	55.4	68.9	80.4	63.7

the replay trajectory improves from 69.0 after CF to 73.3 after GDPO and outperforms the no-replay trajectory on all five benchmarks. The largest final differences are observed on Pick-a-Pic (+11.1) and EqBen (+5.3).

This reversal suggests that replay plays different roles across the two training stages: it is less important for short-term counterfactual adaptation, but becomes beneficial during policy optimization, where retaining earlier structured and counterfactual supervision helps avoid degradation while optimizing the new semantic-policy and causal-evidence objectives. Importantly, the two GDPO checkpoints continue from their respective replay and no-replay CF branches; therefore, this ablation establishes the effect of the complete replay versus no-replay training trajectories, rather than isolating replay within Stage 3 alone.

A.4.3 AGGREGATION ABLATION

We evaluate the sensitivity of VISEXMEM-9B to the rule used to aggregate claim-level support scores. Using the same claim predictions, we compare the arithmetic mean with geometric, harmonic, and minimum aggregation. As shown in Table 9, the arithmetic mean performs best on all five evaluated benchmarks, with the geometric mean as the closest alternative. This result supports the arithmetic mean used in the main method, which gives equal weight to all visually verifiable claims and requires no learned or benchmark-specific aggregation parameters.

A.5 BENCHMARK EVALUATION PROTOCOLS

We use the same fixed VISEXMEM inference procedure across all benchmarks. The evaluation set and meta-evaluation statistic for each benchmark are fixed in advance; no VISEXMEM prompt, decision threshold, or aggregation rule is tuned on benchmark test data. When an evaluation statistic itself includes a calibration step, as in pairwise accuracy with tie calibration, this is applied only at meta-evaluation time to the already-computed metric scores and does not modify VISEXMEM inference.

Winoground and EqBen. Winoground (Thrush et al., 2022) and EqBen (Wang et al., 2023a) evaluate fine-grained image–text matching using two matched image–caption pairs per example. Following the evaluation setting used in our comparisons, we use the complete 400-example Winoground set and the 280-example EqBen-Mini subset. Let (I_0, c_0) and (I_1, c_1) denote the two matched pairs and $S(I, c)$ the metric score. The Text condition is satisfied when each image prefers its matched caption,

$$S(I_0, c_0) > S(I_0, c_1) \quad \text{and} \quad S(I_1, c_1) > S(I_1, c_0),$$

whereas the Image condition is satisfied when each caption prefers its matched image,

$$S(I_0, c_0) > S(I_1, c_0) \quad \text{and} \quad S(I_1, c_1) > S(I_0, c_1).$$

The Group condition requires both the Text and Image conditions to hold. We report all three scores where space permits and use Group for the main comparison because it requires all four matching inequalities to be correct.

TIFA160 and GenAI-Bench. For TIFA160 (Hu et al., 2023), we evaluate the 800 image–text pairs formed from 160 prompts and outputs from five text-to-image models, together with their 1–5 human alignment ratings. For GenAI-Bench (Li et al., 2024a), we use the 527-prompt image coreset used for human and automatic meta-evaluation, rather than the full 1,600-prompt collection.

For both benchmarks, we report pairwise accuracy (PAcc) with tie calibration. PAcc is computed from the vectors of human ratings and metric scores by comparing their induced pairwise higher/lower/tie decisions; the metric's tie threshold is calibrated according to the pairwise-accuracy protocol. Thus, these benchmarks contain scalar human ratings rather than predefined binary preference pairs, and PAcc measures agreement between the pairwise orderings induced by the human and metric scores. Results are reported on the 0–100 scale.

Pick-a-Pic. For Pick-a-Pic (Kirstain et al., 2023), we follow the filtered evaluation subset used in the VQAScore protocol: 100 prompts, each paired with two generated images, for a total of 200 images. Human-tied examples are excluded from this subset. We report binary preference accuracy, where a comparison is correct when the human-preferred image receives the larger metric score for the corresponding prompt.

CROC$_{\mathrm{hum}}$. CROC$_{\mathrm{hum}}$ (Leiter et al., 2026) evaluates contrastive robustness in four directions obtained by crossing text-based versus image-based evaluation with forward versus inverse evaluation. In text-based evaluation, the image is held fixed while the original and contrastive texts are compared; in image-based evaluation, the text is held fixed while original and contrastive images are compared. Forward evaluation treats the original text–image pair as the matching pair, whereas inverse evaluation treats the contrastive pair as the matching pair. We compute all four direction scores using the benchmark protocol and report their unweighted arithmetic mean as the single CROC$_{\mathrm{hum}}$ score in the main comparison.

T2I-CompBench++. For T2I-CompBench++ (Huang et al., 2025), we use its released human-evaluation image–text examples and report Spearman's rank correlation coefficient (SRCC) between automatic metric scores and the corresponding human alignment judgments on the evaluation subset used in our experiments. Spearman correlation evaluates agreement in rank ordering and is invariant to monotonic rescaling of either score. The identical sample set and correlation computation are used for every method in our comparison.

CountBench. CountBench (Paiss et al., 2023) evaluates zero-shot visual counting for counts from two through ten. For each image, the benchmark constructs nine candidate captions by replacing the count in its caption with each number in $\{\text{two}, \ldots, \text{ten}\}$. We score all nine image–caption combinations using the same metric inference procedure and take the count appearing in the highest-scoring caption as the predicted count. We report exact top-1 counting accuracy.

T2IScoreScore. T2IScoreScore (TS2) (Saxon et al., 2024) meta-evaluates image–text metrics using *semantic error graphs* (SEGs). Each SEG associates a prompt with groups of images containing controlled semantic-error configurations. We use the two principal TS2 statistics reported in the main paper. The ordering score rank_m is the benchmark's Spearman-based measure of whether metric scores correctly order images along walks through an SEG according to semantic-error severity. The separation score sep_m measures how well score distributions for neighboring error nodes are distinguished using the two-sample Kolmogorov–Smirnov statistic. Higher values indicate better performance for both measures.

We report rank_m and sep_m over all evaluated SEGs (*All*) and separately for the *Synth*, *Nat*, and *Real* partitions. *Synth* uses generated images to populate predefined semantic error graphs, *Nat* constructs error graphs from natural images, and *Real* contains naturally occurring semantic errors from text-to-image generations. Here, "Real" refers to the errors being naturally occurring rather than to the images being real photographs. Reporting both ordering and separation distinguishes metrics that recover the intended semantic ordering from those that also assign clearly separated scores to different error populations.

AGITA-Bench. For AGITA-Bench (Sec. 4), evaluation is performed separately at the global and atomic levels. Global agreement is Spearman's ρ between each method's image–text score and the human global-alignment rating. Atomic agreement is computed only for methods that can directly score the fixed benchmark claims and is Spearman's ρ between claim-level metric scores and human atomic-support judgments; claims marked *cannot judge* are excluded. Correlations are reported separately for the Easy, Medium, and Hard semantic-complexity subsets, with the reported average

equal to the unweighted mean of the three subset correlations. The benchmark construction and full human-annotation protocol are provided in Appendix A.6.

GroundingDINO baseline. To test how much image-text alignment can be recovered from grounding alone, we replace the VISEXMEM verifier with GroundingDINO-base (Liu et al., 2024). Captions are decomposed into atomic claims by the same frozen text-only VISEXMEM v2 decomposer, which receives neither images nor benchmark labels. Each claim is converted to a lowercase, period-terminated query, and GroundingDINO is the only component that receives image pixels. We use a box threshold of 0.30, text threshold of 0.25, NMS IoU of 0.5, and at most 32 candidate boxes per query. Claim support is based on the strongest phrase-aligned confidence before thresholding, with the scoring rule determined by claim type.

Object claims use the maximum confidence and negated claims use its complement. Count claims combine the strongest detections with agreement between the detected and expected counts. Attribute claims combine entity and attribute confidence with box overlap, while spatial claims also use the relative positions of subject and object boxes. Verb relations combine subject, object, and phrase confidence, and unary actions combine subject and phrase confidence. Depth relations and abstract claims are marked ungroundable because no reliable 2D box rule is defined. If M of the N claims have a defined score, the final metric is

$$S_{\mathrm{GDINO}} = \frac{M}{N} \exp\left(\frac{1}{M} \sum_{i=1}^{M} \log \max(s_i, 10^{-6}) \right).$$

The geometric mean makes low-support claims reduce the overall score, while the coverage term M/N penalizes captions for which only part of the content can be grounded. All thresholds and constants are fixed before evaluation. The baseline uses no training, calibration, or benchmark-specific tuning, and GroundingDINO confidence is treated as an uncalibrated grounding support score rather than a probability.

A.5.1 VLM-AS-A-JUDGE EVALUATION

We evaluate untuned Qwen3.5-4B, Qwen3.5-9B, and GPT-5.6 Sol as VLM-as-a-Judge controls under two fixed inference settings. Results on the eight established benchmarks and AGITA-Bench are reported in Tables 1 and 2, respectively.

Holistic (Hol.). The VLM receives the image and complete caption and directly assigns a single alignment score using a fixed rubric from 0.05 to 0.95. No decomposition, claim-level verification, or localized evidence is used. The complete prompt is shown in Figure 8.

Decomposed (Dec.). The same backbone is evaluated with a structured end-to-end decompose–verify–aggregate procedure. The VLM receives the image and complete caption and, in a single response, decomposes the caption into visually verifiable atomic claims, localizes claim-level visual evidence, assigns each claim a support score in $[0, 1]$, and generates a short evidence-grounded explanation. Possible claim types include *object*, *attribute*, *count*, *action*, *relation*, *spatial*, *negation*, and *other*. After evaluating all claims, the VLM produces an overall score in $[0, 1]$ based on the completed part-level judgments. The overall score is generated by the VLM rather than computed using an external aggregation formula. The complete prompt is shown in Figure 9.

These controls use no VISEXMEM-specific fine-tuning and do not use VISEXMEM's learned evidence-conditioned verifier or external grounding cues. Thus, within each backbone, Hol. versus Dec. changes the inference structure while keeping the underlying VLM fixed.

A.5.2 BASELINE COVERAGE AND REPRODUCIBILITY

REVEALER (Shi et al., 2026) and T2I-FineEval (Hosseini et al., 2025) are particularly relevant to VISEXMEM because both perform fine-grained image–text evaluation rather than relying only on a global similarity score. We therefore examined whether they could be included in the controlled benchmark comparison.

REVEALER. REVEALER performs element-level alignment evaluation through a structured grounding–reasoning–conclusion trajectory and trains its evaluator with reinforcement learning over grounding and alignment objectives. At the time of our evaluation, we were unable to obtain a usable public implementation and released checkpoint sufficient to reproduce the published method on our benchmark suite. Reimplementing the method from the paper would introduce uncontrolled choices in model initialization, training data, reward construction, and inference, and would therefore not constitute a faithful baseline reproduction. We consequently do not report reconstructed REVEALER numbers.

T2I-FineEval. T2I-FineEval decomposes both modalities: its reported configuration uses GPT-4 with two-shot prompting to decompose the input text and generate entity- and relation-level questions, YOLOv9 to obtain image regions, and BLIP-VQA for question–region scoring. Although evaluation code is publicly available, faithfully reproducing the published configuration therefore requires the proprietary GPT-4 decomposition stage and its corresponding prompting configuration. Replacing this stage with a different language model would change the evaluated method and confound the comparison with backbone and prompting differences. Because we could not reproduce the published configuration exactly under our controlled evaluation setup, we do not report a modified T2I-FineEval result.

We also avoid copying numbers reported in the original papers into our main tables, since their evaluated subsets, scoring statistics, and/or underlying models do not match our fixed protocols. All quantitative comparisons in Tables 1 and 2 therefore use scores obtained under the evaluation settings specified in this paper.

VLM-as-a-Judge (Holistic)

```
----- HOLISTIC system -----
You are a strict visual alignment evaluator.
You judge how well a caption matches an image, using only visible
image evidence.

Do not assume something is true because it sounds plausible.
Do not use world knowledge when the image does not show it.
Do not reward fluency or aesthetic quality.

Return only the required JSON schema.

----- HOLISTIC user (image + full caption; no decomposition) -----
Judge how well the caption below matches the supplied image.

Consider every visually verifiable element the caption expresses:
objects, attributes, actions, relations, counts, and negation.
A meaningful contradiction must lower the score even when most of the
caption is correct.

alignment_score rubric (from 0.05 to 0.95):
  0.90-0.95  clearly and directly supported by visible evidence
  0.75-0.85  strongly supported, with minor visual ambiguity
  0.60-0.70  more supported than not, but incomplete/weak evidence
  0.45-0.55  genuinely uncertain / ambiguous / not enough visible
  evidence
  0.30-0.40  more unsupported than supported
  0.15-0.25  strongly unsupported / likely contradicted
  0.05-0.10  clearly contradicted, or required visual evidence is
  clearly absent

Return valid JSON only, in exactly this shape:
{"alignment_score":0.95}

Caption: {caption}
```

Figure 8: Prompt used for the holistic VLM-as-a-Judge setting. The model directly evaluates the complete image–caption pair and returns a single alignment score without atomic decomposition.

VLM-as-a-Judge (Decomposed, End-to-End)

```
You are an expert visual evaluator.

Your task is to evaluate how well the
provided image supports the provided
caption.

You must perform the COMPLETE evaluation
in this single response.

CAPTION:
{caption}

Follow these steps internally in order:

1. Decompose the caption into visually
verifiable atomic claims.
2. For every atomic claim, locate the
visual evidence needed to judge it.
3. Assign an evidence-grounded support
score to every atomic claim.
4. Give one short explanation for each
score based on the visible evidence.
5. Only after evaluating all atomic
claims, produce one overall
   image-caption support score based on
   those part-level judgments.

Do not skip any visually meaningful
component of the caption.

--------------------------------
ATOMIC DECOMPOSITION
--------------------------------

Split the caption into the smallest
useful independently verifiable visual
claims.

Possible claim types include:

- object
- attribute
- count
- action
- relation
- spatial
- negation
- other

Examples of information that should
normally become separate claims when
present:

- an object exists;
- an object has an attribute;
- a specific number of objects exists;
- an entity performs an action;
- one entity acts on another;
- two entities have a relation;
- an entity is left/right/above/below/i
n-front-of/behind another;
- a stated object is absent or a
condition is negated.

Preserve binding.

For example, do not reduce:

"the man in a red shirt is holding a
blue umbrella"

to unrelated statements such as:

"red exists"
"blue exists"
"man exists"
"umbrella exists"

The atomic claims must preserve which
property/action/relation belongs to
which entity.

Do not invent claims not contained in
the caption.

Do not evaluate non-visual linguistic
details that cannot reasonably be
verified from the image.

--------------------------------
BOUNDING BOXES / VISUAL EVIDENCE
--------------------------------

For each atomic claim identify only the
visual entities or regions needed
to verify that claim.

Represent every bounding box using
normalized coordinates:

[x_min, y_min, x_max, y_max]

with each coordinate an integer from 0
to 1000.

Coordinate system:

- (0,0) = top-left of the image
- (1000,1000) = bottom-right of the image

Each evidence item must contain:

- label
- role
- box

Example:

{
  "label": "dog",
  "role": "subject",
  "box": [120, 310, 430, 820]
}

Use multiple boxes when necessary.

Examples:

COUNT:
box all clearly visible relevant
instances needed to judge the count.

RELATION / SPATIAL:
box all relevant participating entities.

ACTION:
box the actor and other important
participant/object where applicable.

ATTRIBUTE:
box the entity to which the attribute is
bound.

INCORRECT CLAIM:
show the visible counter-evidence when
possible.

For example, if the caption claims three
dogs but only two are visible,
return the boxes of those two visible
dogs.

NEGATION / ABSENCE:
if the claim is supported specifically
because something is absent and
there is no meaningful region to box,
`evidence_boxes` may be an empty
list.

Do not invent bounding boxes merely to
satisfy the schema.

--------------------------------
PART-LEVEL SUPPORT SCORE
--------------------------------

For every atomic claim output a support
score in:

[0.0, 1.0]

Interpretation:

1.00
= clearly and fully supported by the
image

0.75
= mostly supported, with minor ambiguity
or uncertainty

0.50
= genuinely ambiguous, partially
supported, or visually uncertain

0.25
= mostly unsupported

0.00
= clearly contradicted or absent

Use the full continuous range when
appropriate.

The score must measure SUPPORT FOR THE
SPECIFIC ATOMIC CLAIM, not generic
image-caption similarity.

Pay particular attention to:

- entity identity;
- attribute binding;
- subject/object roles;
- action participants;
- spatial ordering;
- counts;
- negation;
- same-class entity binding.

Do not give a high score merely because
all relevant words or objects
appear somewhere in the image.

--------------------------------
SHORT EXPLANATION
--------------------------------

For each atomic claim provide ONE
concise evidence-grounded sentence.

The explanation should state the
visually relevant fact that justifies
the
score.

Good examples:

"Two dogs are visible, while the claim
requires three."

"The red shirt is worn by the man,
matching the claim."

"The dog is to the right of the cat, not
to its left."

"The woman is holding the umbrella
rather than the man."

Avoid:

- generic descriptions;
- unnecessary prose;
- hidden chain-of-thought;
- repeating the numerical score;
- mentioning coordinate values in the
explanation.

The explanation must agree with both the
score and the returned evidence
boxes.

--------------------------------
OVERALL SCORE
--------------------------------

After ALL atomic parts have been
evaluated, return:

overall_score

in:

[0.0, 1.0]

The overall score must be derived from
your completed part-level
judgments.

IMPORTANT:

Do NOT independently ignore the part
scores and make an unrelated holistic
judgment.

The overall score should reflect how
well the caption as a whole is
supported by the image given the
atomic-part scores.

A strongly contradicted essential claim
--- especially an incorrect count,
binding, action, relation, spatial
relation, or negation --- should
meaningfully lower the overall score.

At the same time, do not invent an
external formula or expose
chain-of-thought.

The final overall score is YOUR
end-to-end assessment based on the
part-level results you produced above.

--------------------------------
OUTPUT FORMAT
--------------------------------

Return ONLY valid JSON.

No Markdown.
No code fences.
No text before or after the JSON.

Use exactly this top-level structure:

{
  "atomic_parts": [
    {
      "id": "p1",
      "claim": "atomic visual claim",
      "type": "object",
      "evidence_boxes": [
        {
          "label": "entity name",
          "role": "target",
          "box": [0, 0, 1000, 1000]
        }
      ],
      "score": 0.95,
      "explanation": "One short
      evidence-grounded sentence."
    }
  ],
  "overall_score": 0.90
}

Requirements:

- `atomic_parts` must contain at least
one item.
- IDs must be p1, p2, p3, ... in order.
- Every claim must have exactly one
score.
- Every claim must have exactly one
explanation.
- Every score must be within [0,1].
- Every box coordinate must be within
[0,1000].
- `evidence_boxes` may be [] only when
no meaningful localized visual
  evidence exists.
- `overall_score` must be within [0,1].
- `overall_score` must be produced after
considering all atomic-part
  judgments.

Return only the JSON object.
```

Figure 9: Prompt used for the decomposed end-to-end VLM-as-a-Judge setting. The model performs atomic decomposition, evidence localization, claim-level scoring, short explanation generation, and final overall scoring in a single response.

A.6 AGITA-BENCH

A.6.1 BENCHMARK CONSTRUCTION

AGITA-Bench contains 600 image-text pairs, split equally between I2T and T2I evaluation. The 300 I2T cases use 100 real images each from ADE20K (Zhou et al., 2017), HICO-DET (Chao et al., 2018), and Objects365 (Shao et al., 2019). Their captions cover high, partial, and low alignment. The 300 T2I cases use original generation prompts paired with 181 images from DiffusionDB (Wang et al., 2023b) and 119 from MJHQ-30K (Li et al., 2024b).

We construct 200 easy, 200 medium, and 200 hard cases based on *semantic complexity* rather than alignment severity. Fixed rules assign difficulty using the number and composition of atomic claims. Easy cases contain few simple claims. Medium cases add relations or attribute and count assignments. Hard cases contain more constraints or require correct binding of attributes, counts, roles, or spatial relations. The resulting benchmark contains 8,735 atomic claims. AGITA-Bench is excluded from training, calibration, and model selection.

Atomic decomposition prompt. AGITA-Bench captions are decomposed into typed visual semantic units using GPT-5.6 Sol with a fixed caption-only prompt. The model does not receive the image during decomposition. It is instructed to preserve entity and attribute bindings, counts, actions, relations, spatial relations, and negation while avoiding unobservable or stylistic content. The complete prompt is shown in Figure 10.

A.6.2 HUMAN ANNOTATION PROTOCOL

AGITA-Bench is annotated at two complementary levels. For each atomic claim, annotators assign an image-support score from 0, meaning contradicted or not supported, to 4, meaning fully supported. A separate *cannot judge* option is available when the visual evidence is insufficient for a reliable decision. Each image-text pair is also rated for global alignment from 1, meaning not aligned, to 5, meaning fully aligned. Decomposition quality is judged independently of image support. A claim is *reasonable as written* when it is a clear and non-redundant atomic proposition that requires neither splitting nor merging. This separates decomposition errors from genuine image-text mismatches.

AGITA-Bench provides human score annotations rather than human bounding-box annotations. Any localized evidence boxes reported by VISEXMEM are produced by the metric and are not part of the benchmark's human labels.

All seven annotators independently annotate the same 30 image-text pairs, comprising 460 unique atomic claims, for inter-annotator agreement analysis. The remaining 570 pairs are each assigned to one annotator. Pairwise quadratic-weighted Cohen's κ is reported in Tables 10 and 11. Aggregate annotator demographics are reported in Table 12.

A.6.3 ANNOTATION INTERFACE, GUIDELINES AND EXAMPLES

Annotations are collected through a dedicated interface that presents the same information and rating controls to every annotator. Figure 11 shows the dashboard and sample-level interface. Before starting, annotators can review the annotation guide and worked examples. For each assigned sample, they record decomposition quality and atomic-support judgments for individual claims alongside the image and input text. Global image-text alignment is rated once for the complete sample. The full annotator instructions are shown in Figure 12.

We also provide four worked examples covering high, partial, and weak image-text alignment, including one example that shows the difference between decomposition quality and visual support. Each example includes the input, global alignment judgment, and claim-level decomposition and support annotations.

Table 10: **Inter-annotator agreement for global alignment.** Pairwise quadratic-weighted Cohen's κ on the 30 shared image–text pairs; mean pairwise $\kappa = .602$. Only unique annotator pairs are shown.

	Ann2	Ann3	Ann4	Ann5	Ann6	Ann7
Ann1	.57	.78	.54	.76	.37	.62
Ann2		.62	.70	.64	.54	.48
Ann3			.73	.81	.39	.62
Ann4				.63	.47	.61
Ann5					.51	.65
Ann6						.61

Table 11: **Inter-annotator agreement for atomic support.** Pairwise quadratic-weighted Cohen's κ on shared claims; *cannot judge* responses are excluded pairwise ($N = 418$–453). Mean pairwise $\kappa = .564$.

	Ann2	Ann3	Ann4	Ann5	Ann6	Ann7
Ann1	.52	.72	.61	.60	.47	.39
Ann2		.59	.60	.58	.59	.55
Ann3			.65	.58	.67	.47
Ann4				.55	.55	.49
Ann5					.56	.45
Ann6						.66

Table 12: **AGITA-Bench annotator demographics.** Demographic summary of the seven human annotators ($N = 7$).

Attribute	Category	Count
Nationality	German	3
	Indian	1
	Chinese	1
	UAE	2
Education	Postdoctoral researcher	1
	PhD student	2
	Master's student	4
Age	18–25	4
	26–30	2
	> 30	1

AGITA-Bench Atomic Decomposition

```
You split a caption into VISUAL SEMANTIC UNITS for
a compositional benchmark.

You receive ONLY the caption text. You have not
seen any image. Never add anything the caption does
not state, and never judge whether it is true.

THE UNITS ARE PHRASES, NOT SENTENCES.

Stay close to the caption's own wording. Do not
rewrite meaning into full propositions. This is the
single most important rule.

caption: "A person is moving green hay towards an
elephant that is inside the back of a white truck."

  CORRECT
  type
    person
    object
    hay
    object
    green hay
    attribute
    elephant
    object
    truck
    object
    white truck
    attribute
    moving
    action
    green hay towards an elephant
    relation
    elephant inside the back of a white truck
    spatial

  WRONG -- do not produce this
    A person exists.
    Hay exists.
    The hay is green.
    An elephant exists.
    A truck exists.
    The truck is white.

caption: "Dozens of mostly red apples hang from the
branches of a leafy green fruit tree."

  CORRECT
    apples
    object
    dozens of apples
    count
    mostly red apples
    attribute
    branches
    object
    fruit tree
    object
    leafy fruit tree
    attribute
    green fruit tree
    attribute
    hang
    action
    apples hang from the branches of a fruit tree
    relation

  WRONG
    There are dozens of apples.
    Most of the apples are red.
    The fruit tree is leafy.
    The fruit tree is green.

BINDING IS THE POINT. An attribute, count or role
must stay attached to the entity the caption gives
it.

  "The woman wears blue and the man wears red."
     woman / man / woman wears blue / man wears red
     NOT: woman / man / blue / red

  "The dog chases the cat." -> dog chases cat
     (dog is the agent; never "cat chases dog")

  "one black dog and one white dog" ->
     dogs / black dog / white dog

  "A red cup is to the left of a blue bowl." ->
     red cup / blue bowl / red cup left of blue
     bowl

  "A man without a hat" ->
     man / man without a hat
     (keep the negation)

TYPES -- use exactly these:

  object
    a visible entity, as a compact phrase:
    person, dog, fruit tree, branches

  attribute
    a property bound to its entity:
    green hay, white truck, mostly red apples

  count
    a quantity bound to its entity:
    three dogs, dozens of apples

  action
    what is happening, compactly:
    moving, hang, chases, holding

  relation
    who/what relates to whom/what:
    apples hang from the branches

  spatial
    an arrangement, direction preserved:
    cup left of bowl, cat under table

  negation
    an absence or denial:
    no dog, man without a hat

  other
    visually checkable, none of the above

GRANULARITY -- the smallest USEFUL unit, not every
lexical combination.

  "leafy green fruit tree" ->
     fruit tree, leafy fruit tree, green fruit tree

  NOT also:
     tree, leafy tree, green tree,
     leafy green tree, leafy green fruit tree

  But "dog" AND "black dog" are BOTH wanted:
     presence and attribute are different checks.

  Likewise "dogs" AND "three dogs".

OMIT entirely: style, medium, quality and render
words
(8k, cinematic, unreal engine, masterpiece,
trending on artstation,
photorealistic, octane render, artist names),
camera and lens jargon,
aspect ratios, mood words with no visual test
(epic, dreamy), and
unobservable facts about a subject (their
nationality, their profession
when not visible).

Return STRICT JSON, no prose:

{"parts": [{"text": "...", "type": "object"},
           {"text": "...", "type": "attribute"}]}

Do not number the parts and do not invent ids.
```

Figure 10: Prompt used for AGITA-Bench atomic decomposition. GPT-5.6 Sol receives only the caption and decomposes it into typed visual semantic units without access to the image or making any image-based correctness judgment.

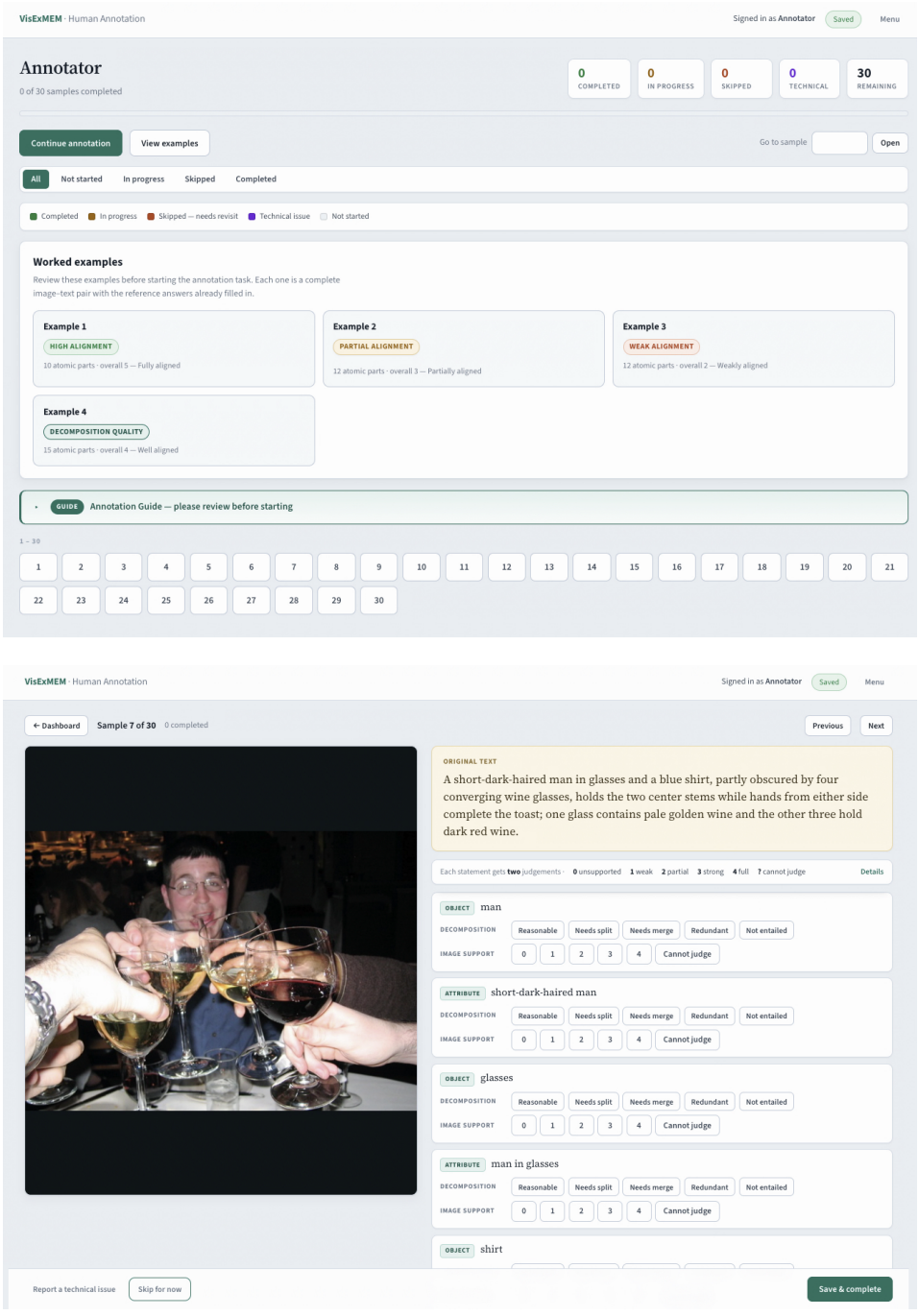

Figure 11: **AGITA-Bench annotation interface.** Dashboard and sample-level interface used for human annotation.

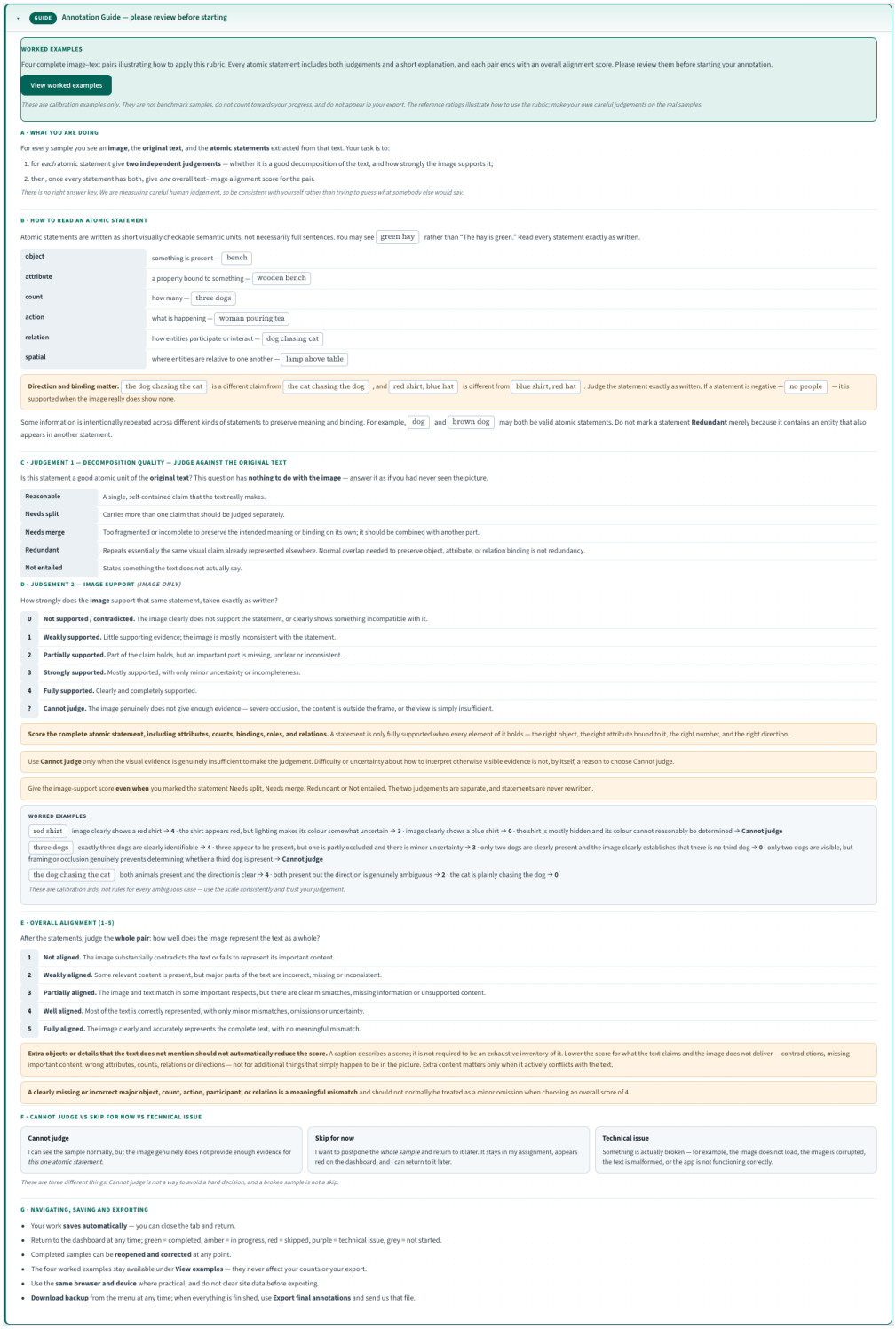

Figure 12: **AGITA-Bench annotation guidelines.** Complete annotator-facing guidelines.

Example 1: High alignment

High alignment



Input text

A man works on a red bicycle in a workshop with tools hanging on a green pegboard.

Global alignment score 5 / 5

#	Type	Atomic claim	Decomp. quality	Atomic support
1	object	man	Reasonable	**4**
2	object	bicycle	Reasonable	**4**
3	attribute	red bicycle	Reasonable	**4**
4	object	workshop	Reasonable	**4**
5	object	tools	Reasonable	**4**
6	object	pegboard	Reasonable	**4**
7	attribute	green pegboard	Reasonable	**4**
8	action	man works on a red bicycle	Reasonable	**4**
9	spatial	man in a workshop	Reasonable	**4**
10	relation	tools hanging on a green pegboard	Reasonable	**4**

Overall rationale

The image clearly supports the man, red bicycle, workshop, green pegboard, tools, and the described actions and relations.

Example 2: Partial alignment

Partial alignment



Input text

A black dog lies on a patterned sofa beside several pillows, with a red ball near its front paws.

Global alignment score 3 / 5

#	Type	Atomic claim	Decomp. quality	Atomic support
1	object	dog	Reasonable	**4**
2	attribute	black dog	Reasonable	**0**
3	object	sofa	Reasonable	**4**
4	attribute	patterned sofa	Reasonable	**4**
5	object	pillows	Reasonable	**4**
6	count	several pillows	Reasonable	**4**
7	object	ball	Reasonable	**0**
8	attribute	red ball	Reasonable	**0**
9	object	front paws	Reasonable	Cannot judge
10	action	dog lies on a patterned sofa	Reasonable	**4**
11	spatial	dog beside several pillows	Reasonable	**4**
12	spatial	red ball near the dog's front paws	Reasonable	**0**

Annotation note. **p9 (front paws):** The dog's front paws are sufficiently obscured that this standalone part cannot be confidently judged from visible evidence.

Overall rationale

The dog, patterned sofa, pillows, and lying position are well supported, but the dog is brown rather than black and no red ball is visible.

Example 3: Weak alignment

Weak alignment



Input text

A Shiba Inu wearing a red beret and a blue turtleneck sits in front of a green wooden wall.

Global alignment score 2 / 5

#	Type	Atomic claim	Decomp. quality	Atomic support
1	object	Shiba Inu	Reasonable	**4**
2	object	beret	Reasonable	**4**
3	attribute	red beret	Reasonable	**0**
4	object	turtleneck	Reasonable	**4**
5	attribute	blue turtleneck	Reasonable	**0**
6	object	wall	Reasonable	**4**
7	attribute	green wall	Reasonable	**0**
8	attribute	wooden wall	Reasonable	**0**
9	relation	Shiba Inu wearing a red beret	Reasonable	**2**
10	relation	Shiba Inu wearing a blue turtleneck	Reasonable	**2**
11	action	Shiba Inu sits	Reasonable	Cannot judge
12	spatial	Shiba Inu in front of a wall	Reasonable	**4**

Annotation notes.

p9 (Shiba Inu wearing a red beret): Partially supported: the dog IS wearing a beret, but the required red attribute is wrong. A composite relation that is partly correct but contains an important mismatch scores 2, not 0.

p10 (Shiba Inu wearing a blue turtleneck): Partially supported for the same reason: the turtleneck is present but black rather than blue.

p11 (Shiba Inu sits): The lower body is outside the frame, so the sitting posture cannot be judged confidently.

Overall rationale

The Shiba Inu, beret, turtleneck, and wall are present, but the beret is black rather than red, the turtleneck is black rather than blue, and the wall is light-colored brick rather than green wood. The lower body is outside the frame, so the sitting posture cannot be judged confidently.

Example 4: Decomposition quality

Decomposition quality

Teaching point. Decomposition quality and image support are independent: this example pairs every decomposition label with a high support score, and a Reasonable decomposition with a support score of 0.



Input text

A woman in a red shirt is holding a blue umbrella beside a brown dog, while standing on a wet sidewalk in front of a yellow bus.

Global alignment score **4 / 5**

#	Type	Atomic claim	Decomp. quality	Atomic support
1	object	woman	Reasonable	**4**
2	attribute	red shirt	Reasonable	**4**
3	attribute	blue umbrella	Reasonable	**4**
4	attribute	brown dog	Reasonable	**4**
5	attribute	wet sidewalk	Reasonable	**4**
6	object	bus	Reasonable	**4**
7	attribute	yellow bus	Reasonable	**0**
8	relation	woman holding a blue umbrella	Reasonable	**4**
9	spatial	woman beside a brown dog	Reasonable	**4**
10	relation	woman standing on a wet sidewalk	Reasonable	**4**
11	spatial	woman in front of a yellow bus	Reasonable	**2**
12	relation	woman in a red shirt holding a blue umbrella beside a brown dog	**Needs split**	**4**
13	action	holding	**Needs merge**	**4**
14	attribute	shirt that is red	**Redundant**	**4**
15	action	smiling woman	**Not entailed**	**4**

Annotation notes.

p7 (yellow bus): The text correctly decomposes to ‘yellow bus’, but the visible bus is red.

p11 (woman in front of a yellow bus): The woman is in front of a bus, but the required yellow attribute is incorrect.

p12 (woman in a red shirt holding a blue umbrella beside a brown dog): It combines several independently checkable claims, although the complete statement is visually supported.

p13 (holding): The action is visually present, but as a decomposition it loses who is holding what.

p14 (shirt that is red): It repeats the visual claim already represented by ‘red shirt’.

p15 (smiling woman): The original text never says the woman is smiling, but the image clearly shows her smiling.

Overall rationale

The woman, red shirt, blue umbrella, brown dog, wet sidewalk, and described actions and spatial relations are clearly supported. The meaningful mismatch is the bus colour: the text specifies a yellow bus, while the image shows a red bus.

Table 13: **Inter-annotator agreement for explanation and evidence evaluation.** Values show mean pairwise agreement. Ordinal ratings use quadratic-weighted Cohen's κ, while *Unsupported Content* uses unweighted κ.

Evaluation	**Mean κ**	**Within ± 1 (%)**
Task A explanation quality	.59	89.5
Unsupported Content	.42	–
Task B localized evidence	.58	86.9

A.7 Explanation and Evidence Annotation

Protocol. Each sample presents four explanations from EXPERT, VIEScore, GPT-5.6 Sol, and VisExMem. System identities are hidden and the explanations are shown under randomized labels *A/B/C/D*. Annotation has two stages. In Task A, annotators rate all four explanations without seeing localized evidence. They assess diagnostic correctness, visual grounding, clarity, and diagnostic usefulness on a 1 to 5 scale, and flag unsupported factual content. For systems that provide localized evidence, Task B is shown only after Task A is complete. Annotators then rate up to three claim-region pairs for localization accuracy and evidence sufficiency, again on a 1 to 5 scale. Task A responses are locked before Task B is shown so the highlighted regions cannot affect the earlier explanation ratings. The complete questions, scales, and instructions are reproduced in Figure 13.

Annotators and agreement. Three annotators took part in the study. Their reported nationalities were German, Indian, and UAE. Two were aged 18 to 25 and one was aged 26 to 30, and all three were Master's students. Agreement is computed pairwise on items completed by both annotators in each pair after resolving the randomized labels to their underlying systems. For the ordinal ratings, we use quadratic-weighted Cohen's κ, while *Unsupported Content* uses unweighted Cohen's κ. Across the four explanation-quality criteria, mean pairwise weighted κ is .59, with 89.5% of paired ratings within one scale point. *Unsupported Content* gives $\kappa = .42$. For localized evidence, mean weighted κ is .58, with 86.9% within-one-point agreement. Because the evidence ratings are strongly concentrated near the top of the scale, we report near-agreement together with κ.

HUMAN EXPLANATION & EVIDENCE ANNOTATION GUIDELINES

A · WHAT YOU ARE DOING

For every sample you see an **image** and the **caption** being evaluated. Four different automatic systems each wrote a short explanation of whether the image supports the caption. The explanations are shown only as **Explanation A / B / C / D**; the system identity is hidden and the order varies across samples.

1. **Task A** — rate each of the four explanations, one at a time. No highlighted regions are shown yet.
2. **Task B** — after all four explanations are rated, a second panel opens only for explanation(s) whose system provides a highlighted image region for a claim. Depending on the sample, this may apply to none, one, or multiple explanations. Up to three highlighted regions are rated. Once Task B is revealed, the corresponding **Task A answers are locked and can no longer be changed**, keeping the two judgments independent.

Not every explanation has a Task B. Some systems do not produce localized image regions; this is expected and should not be flagged.

B · TASK A — RATING ONE EXPLANATION

Diagnostic Correctness	"How accurately does the explanation identify the relevant matches or mismatches between the image and caption?" *1 Completely incorrect · 2 Mostly incorrect · 3 Partially correct · 4 Mostly correct · 5 Completely correct*
Visual Grounding	"How well are the statements in the explanation supported by what is visibly present in the image and stated in the caption?" *1 Not supported · 2 Mostly unsupported · 3 Partially supported · 4 Mostly supported · 5 Fully supported*
Clarity	"How easy is the explanation to understand, without unnecessary or confusing information?" *1 Very unclear · 2 Somewhat unclear · 3 Neither clear nor unclear · 4 Clear · 5 Very clear*
Diagnostic Usefulness	**Primary question.** "How useful is the explanation for understanding why the image and caption align or fail to align?" *1 Not useful · 2 Slightly useful · 3 Moderately useful · 4 Very useful · 5 Extremely useful*
Unsupported Content	"Does this explanation contain any specific factual statement that is contradicted by, or unsupported by, the image/caption?" *Yes / No / Unsure*

Judge every explanation against the image yourself.

C · TASK B — RATING HIGHLIGHTED EVIDENCE

For each qualifying explanation, annotators evaluate up to three claims, each accompanied by its own highlighted image region and short claim text.

Evidence Localization Accuracy	"How accurately do the highlighted region(s) identify the visual evidence relevant to this claim?" *1 Completely incorrect · 2 Mostly incorrect · 3 Partially correct · 4 Mostly correct · 5 Completely correct*
Evidence Sufficiency	"How sufficient are the highlighted region(s) for deciding whether this claim is supported by the image?" *1 Completely insufficient · 2 Mostly insufficient · 3 Partially sufficient · 4 Mostly sufficient · 5 Fully sufficient*

D · SKIP FOR NOW / TECHNICAL ISSUE

Skip for now	Postpone the whole sample. It remains in the annotator's assignment and is marked as unfinished on the dashboard until completed.
Technical issue	Use only when something is actually broken, such as an image failing to load, garbled text, or the annotation interface malfunctioning.

E · NAVIGATING, SAVING AND EXPORTING

- Work **saves automatically**; annotators may close the tab and return later.
- Completed samples can be **reopened**, but after Task B has been revealed for a sample, its Task A answers remain locked. Task B ratings may still be revised.
- Annotators are asked to use the **same browser and device** where practical and not clear site data before exporting.
- A **backup** can be downloaded from the menu at any time. After completing all assigned samples, annotators use **Export final annotations** to produce the final annotation file.

Annotation order

Image + Caption ⟶ **Task A: Explanation** ⟶ **Lock Task A** ⟶ **Task B: Evidence**

Figure 13: **Human explanation and evidence annotation guidelines.** Complete instructions provided to annotators.