Datasets for Training and Evaluating Image-Editing Judges: Preference Data, Label Provenance, and Failure Modes

Joris Strakeljahn, Viktoriia Vovchenko, Sergej Schultenkämper, Frederik S. Bäumer
Bielefeld University of Applied Sciences and Arts
Bielefeld, Germany
{joris.strakeljahn, viktoriia.vovchenko, sergej.schultenkaemper, frederik.baeumer}@hsbi.de

***Abstract*—Learned judges and reward models for instruction-guided image editing are trained on preference datasets that grew quickly from 2023 to 2026. Papers usually report how many preference pairs they contain, but those counts mix different evidence. For example, EditHF-1M expands averaged rankings of 44.3 thousand groups into 29.1 million pairs, EditReward-Data derives roughly 200 thousand pairs from single-expert ratings, and EditScore reports roughly 70 thousand training samples scored by a model, with no human labels in training. A pair count therefore does not equal the number of independent human decisions behind it. To examine this, we compare sixteen training and test sets that cover the different ways these labels are made. For each set we record where the labels come from and whether a test reuses the training pipeline. For the benchmarks, we also record how accuracy is defined and which error types have human reference labels. From this comparison we derive a shared vocabulary for how labels are made, a table that matches data to a development goal, and a nine-item reporting checklist that says what to publish with each set. The checklist is needed because benchmarks that report the same accuracy often measure different quantities, and because nine of the sixteen sets carry no stated data license, seven of them public. We recommend reporting the raw judgments, the evaluation protocol, and the data license with every accuracy number.**

***Index Terms*—image editing evaluation, preference data, reward models, vision-language judges, label provenance**

I. Introduction

Instruction-guided image editing takes a source image and a written instruction and produces a modified image. A good edit makes the requested change and leaves the rest of the image as it was. The model that scores both of these is called a judge. Similarity scores, such as a directional CLIP score, only approximate them [1], [2]. Judges themselves fall into two kinds. A prompted judge, such as VIEScore, asks a vision-language model for separate scores on a fixed rubric [3]. A learned judge trains a scoring function on labeled scores or preferences. When its score serves as a training reward, we call it a reward model. Both kinds are used in practice: they pick the best of several candidate edits, and they supply the reward when an editor is trained with reinforcement learning [4].

Papers usually report a learned judge by the number of preference pairs in its training set, but that count does not record how the pairs were obtained. EditHF-1M, for instance, expands group rankings into millions of pairs [5], EditReward-Data derives pairs from single-expert ratings [6], and EditScore trains on model scores, with no human labels in training [4]. The path from the judgment to the pair differs in each set, and three further obstacles remain before a reported number can be read across papers:

1) Benchmarks that all report "pairwise accuracy" do not measure the same quantity (Section V). They differ in what is compared, how ties are handled, how the reference labels from human raters are built, and what score chance alone would produce.
2) Separate test sets document separate failure modes (Section VI). They differ in which failures have human reference labels and in how the candidate images are built. The modes include missing a visible defect, preferring an edit that changes more than the instruction asked, and scoring an edit without looking at the source image. That last failure is shown from a judge's attention, not from human labels [7].
3) Some benchmarks share a name, so a reported score cannot be traced to one of them. For example, two benchmarks are called EditReward-Bench, although they use different candidates, labels, and metrics [4], [6]. In this paper we therefore call them Wu-EditReward-Bench and Luo-EditScore-Bench (Section V).

This survey arranges the datasets and benchmarks that train and test these judges. We compare training and test sets, not papers, because one paper often releases several, labeled in different ways. So this survey makes three contributions:

1) *A review of sixteen sets* (Sections III and IV, Fig. 1, Table I): where the labels come from (their provenance), whether a held-out test reuses the training sources, prompts, tasks, and candidate models, and whether the data has its own license, distinct from the license of the code and of the weights.
2) *A comparison of the evaluation protocols* (Section V, Table II): what is compared, how ties are handled, how the human reference labels are built, what the chance score is, and whether pairs were expanded from one ranking.
3) *A map of which error types have human reference labels* (Section VI, Table III), together with a table that matches a development goal to suitable sets (Section VII, Table IV) and a nine-item reporting checklist (Section VIII).

This survey does not rank judges across benchmarks, because the protocols are incompatible. It leaves open whether a judge trained on one set still works on another, because the sets examined here contain no controlled transfer study for editing judges (Section VIII). It also does not certify any benchmark as free of pretraining contamination, because that cannot be checked from the outside for a proprietary model.

II. Scope, Review Boundary, and Related Surveys

A. Inclusion Rule

A set is included in this survey when it contains at least one source image, one edit instruction, one or more candidate edits, and a label that evaluates the edit. It must also be used to train an editing judge or reward model, or to test a judge on human reference labels that the judge's authors did not produce. We call that second use meta-evaluation. The set is the unit because the label pipeline can change inside one paper. Wu et al. [6], for example, release two sets that do not support the same claims: the training set EditReward-Data, labeled by one expert per edit, and the benchmark Wu-EditReward-Bench, labeled by three separate expert groups. A training pair computed from one rating per edit, and a test label decided by a majority of three expert groups, answer different questions. ReasonEdit is the exception: its labels evaluate written critiques rather than the edited images, so Section III treats that as its own label object.

Section IV selects sixteen sets that meet these conditions and lists them in Table I, while a resource that misses a condition is cited only for contrast and is therefore not a row. ImageReward shows why the source image is required: it established learned preference scoring, but it has no source image, so it cannot test whether untouched content was preserved [8]. The same requirement decides the general judge benchmarks, so we leave out those that only score tasks such as image captioning, and we include an editing subset where one exists, as in Multimodal RewardBench 2 (MMRB2) [9]. Editor training sets such as MagicBrush and Emu Edit come in on that same condition, only when a later paper builds an evaluation set on top of them [10], [11], and video and 3D editing stay outside this survey.

B. Corpus and Review Method

We present a semi-systematic survey, following Snyder [12]. It differs from a systematic review in two ways that matter here. First, one reviewer did the screening, and second, the search stopped when a further round of citation walking added no new label path, protocol, or failure mode. We therefore do not claim a complete census. The corpus is English-language arXiv and open-access work from January 2023 to August 2026. We searched Semantic Scholar, arXiv, OpenAlex, and Google Scholar with system names such as "EditReward" and phrases such as "preference image editing", and we followed citations from the papers already found. The resulting records were deduplicated by arXiv identifier and screened with the inclusion rule above. We checked the public releases on GitHub and Hugging Face, rather than copying release claims from the papers. A further round of citation walking checked the claims in Sections IV, V and VIII. All access and license statements refer to the August 2026 snapshot.

C. Related Surveys

This subsection places the survey against the nearest published surveys. Zang et al. [13] survey instruction-based image editing: task definitions, how editor training data are built, editor models, how editors are evaluated, and applications. This survey sits one layer under that. It covers the data that editing judges learn from, and the human reference labels against which the judges themselves are checked. An editor training set matters here only when someone derives an evaluation set from it. In the version of Zang et al. we examined, the judge training sets at the center of this survey, among them EditReward-Data, EditHF-1M, and the EditScore training set, are not treated as their own data category.

Huang et al. [14], Nguyen et al. [15], and Wang et al. [16] survey editing methods and mention evaluation, but they do not analyze judge training data. Zhong et al. [17] survey reward models, with the emphasis on text, so the editing-specific label pipelines in this paper fall outside that coverage. Studies of multimodal models used as judges contribute three protocol names that we reuse: a score for one output, a comparison of two outputs, and a ranking of a batch [18]. So in the English-language corpus we examined, snapshot August 2026, we found no earlier survey that orders image-editing judge data together by where the labels come from, what the data are used for, how the benchmark score is defined, and which failure modes have human reference labels.

III. From Raw Signals to Judge Supervision

The comparisons start here, with the path from a raw judgment to a label, because what a dataset can support depends on how that label was produced. Fig. 1 on the next page traces eight resources along this path, one resource per row, through five stages. The stages are the *raw signal* a person or a model actually produced, the *aggregation* step that merges or filters those signals, the expansion into preference pairs, the *final label unit*, and the *use* of the labels. We call those preference pairs derived pairs. One raw judgment is one response from a rater or a model, so it is one rating of one edit, one pairwise verdict, or one ranking of one group. By contrast, a derived pair is not a further judgment. HIVE shows the difference: one ranking of five edits becomes ten training pairs, and all ten pairs rest on that single ranking. The last stage of this figure is then the use of that label, which is training, an internal test, or an external test. An internal test is a held-out part of the same project's data, and an external test comes from outside that project. Fig. 1 marks them I and E. These are the datasets' own tests, not experiments run for this survey. The paragraphs below follow the rows of Fig. 1, and two of them add a second resource of the same kind (SpatialReward, ADIEE). Pair counts on the right of a row do not measure the annotation effort on its left.

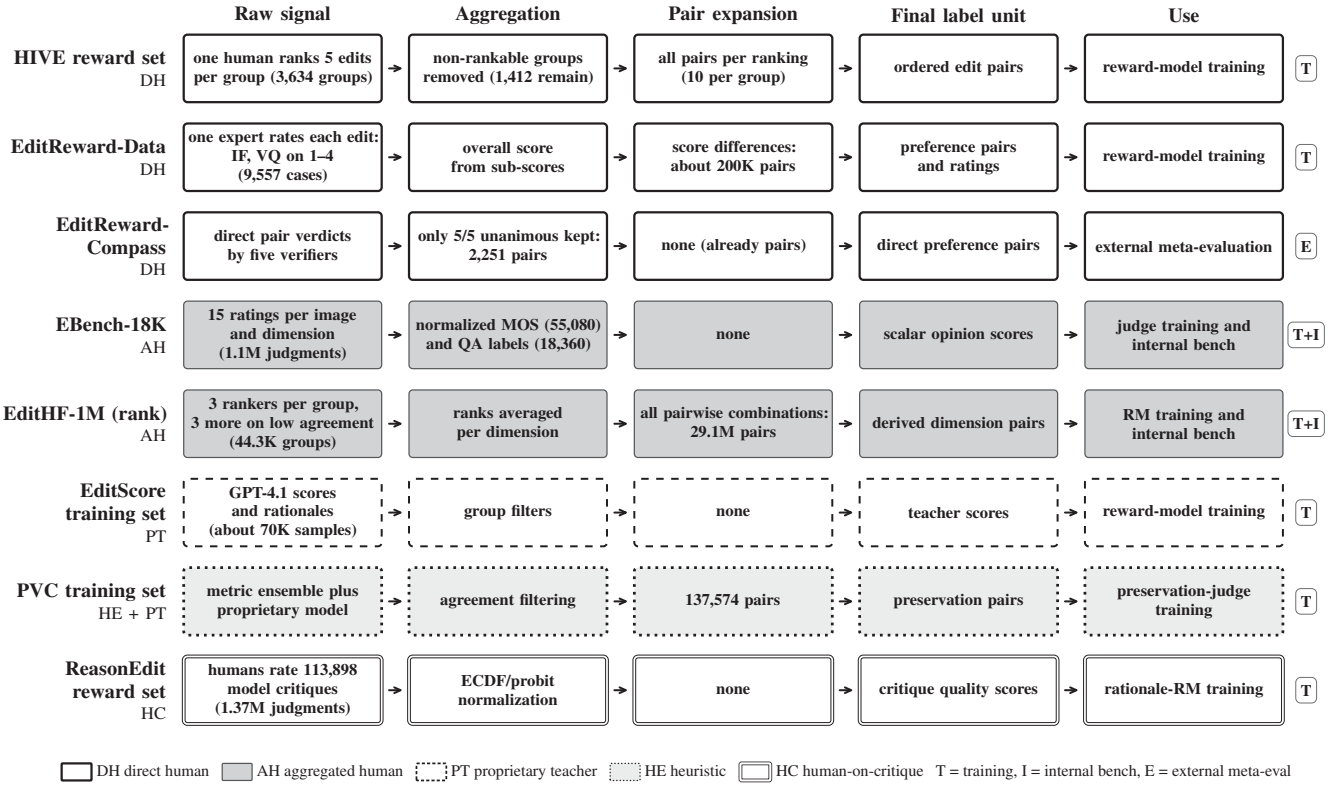


Fig. 1. Eight representative resources traced from raw signal to use. Row style marks provenance (the code under each name, and the legend). The end badge marks use: training (T), internal test (I), external test (E). Pair counts on the right of a pipeline do not measure independent annotation effort on its left. Counts follow the cited papers and releases [4]–[6], [19]–[23].

Direct human rankings. The first row, HIVE [19], is the earliest case in this survey. One annotator ranks five candidate edits per group, and non-rankable groups are removed (1,412 of 3,634 collected groups remain). Each ranking becomes ten ordered pairs for reward-model training. The raw signal and the training labels stay close together, but the cost is a single-rater signal with no agreement estimate.

Human ratings with pair derivation. EditReward-Data [6] collects one expert rating per candidate edit (instruction following and visual quality, each on a 1 to 4 scale) over 9,557 source-instruction cases drawn from six existing editing benchmarks. From the score differences it derives roughly 200 thousand training pairs. The pairs are a computational artifact of the ratings: they were never elicited as pairwise judgments.

Unanimity-filtered direct pairs. EditReward-Compass [21] collects direct pairwise verdicts and keeps only the pairs on which five verifiers agree unanimously, resulting in 2,251 pairs. The size of the candidate pool before the filter is not reported. The result is a unanimity-filtered signal biased toward clearer preferences.

Dense aggregated opinion scores. EBench-18K [20] works at the opposite aggregation extreme: 1,101,600 raw judgments from 30 trained annotators, 15 per image and dimension, are normalized into 55,080 mean opinion scores (MOS) plus 18,360 question-answering labels over 18,360 edits. The label unit is a scalar with a documented rater population, not a preference.

Combinatorial pair expansion. EditHF-1M [5] ranks 44.3 thousand groups covering 1.01 million edits with three professional rankers per group (three more are added when agreement is low). The ranks are averaged per dimension, and every group is expanded into all pairwise combinations, yielding the advertised 29.1 million dimension pairs (pairs counted per rating dimension). These pairs inherit the noise of the averaged rankings and are statistically dependent within each group. Moreover, the paper does not report the exact number of raw judgments behind them.

Proprietary teacher scores. EditScore reports about 70 thousand training samples [4]. GPT-4.1 scores the candidates, and group-level filters are applied while the set is built. The figure of 70 thousand is the paper's count from before that filtering. SpatialReward trains on 260 thousand samples whose scores come from proprietary teacher models that were told where the edit should happen [7]. No person judged the training items, so human evidence enters only at meta-evaluation time.

Heuristic and metric-ensemble labels. The PVC row in Fig. 1 computes the label instead of asking a person for it. The PVC (pairwise visual consistency) training set, the training-side companion of VCReward-Bench in the GEditBench v2 paper, labels 137,574 preservation pairs with a mix of automatic metrics and a proprietary model [22]. ADIEE is the second resource of that kind and scores 93,915 single-turn outputs and 21,382 multi-turn sequences from ground-truth edits and

embedding similarities [24]. The PVC scores themselves were not in the public release we checked. These pipelines scale without annotation cost, but they import the biases of the metrics they use.

Human judgments about critiques. The rows above label the edited image. ReasonEdit [23] labels the written critique instead: 1,366,776 human judgments rate 113,898 model-generated critiques for logicality, accuracy, and usefulness.

Taken together, these pipelines produce labels that look alike in a table header but behave differently in training and evaluation. The 29.1 million pairs expanded from averaged group rankings, the 200 thousand pairs from single-rater scores, and the 70 thousand teacher scores are not the same kind of evidence. They differ in scale, in how independent the labels are, and in how transparent they are. Section IV compares the resources along these axes, but none of the comparisons implies a quality ordering among human, teacher, and heuristic labels.

IV. Judge Datasets by Provenance and Use

Table I covers sixteen sets and can be found on the next page. It is not a ranking by fame, and it is not every set that passed the inclusion rule. Instead, a set becomes a row when it meets at least two of seven criteria and adds something no row yet carries: a historical or scale anchor, a distinct label path, a distinct comparison, an independent meta-evaluation, human labels for a failure mode, a new domain or label object, or a documented train/test split, tie, or overlap risk. These criteria are not a score, so another reader could keep different rows from the same list. A resource that repeats a path already in the table stays in the text without a row, and the sentences below say where it is discussed. ImagenHub and VIEScore anchor Section V. AugCLIP anchors Section VI. The two EditReward-Bench releases are separated in Table II. ADIEE is the computed-label case in Section III, and TIEdit is discussed under access [28]. RealPref-50K carries pseudo-labels from a prompted vision-language model and serves as post-training data for an editor, so it falls outside the inclusion rule [29]. Liu et al. [30] rate one hundred HumanEdit cases from one editing model on twelve factors. Those ratings are discussed with the failure modes (Section VI) and do not add a row.

A. History and Scaling

Table I keeps the raw judgments and the derived pairs in separate columns. From HIVE (1,412 usable rankings, 2023) through EditReward-Data (9,557 rated cases, 2025) to EditHF-1M (1.01 million ranked edits, 2026), the derived counts grow faster than the human decisions behind them. HIVE turns each five-way ranking into ten pairs. EditReward-Data turns per-edit ratings into roughly 200 thousand pairs. EditHF-1M turns averaged group rankings into 29.1 million dimension pairs. Along this line each edit enters more derived pairs, 4 in HIVE, 6 in EditReward-Data, and about 22 per dimension in EditHF-1M (about 657 dimension pairs per ranked group), so the pairs depend more and more on the same human decisions, and a larger pair count is not a larger set of human decisions.

B. Teacher and Heuristic Supervision

The sets in that scaling line still start from a human ranking or a human rating. However, a second group of training sets never asks a person. In that context, a teacher is a model that writes the training scores in place of that person. For EditScore that teacher is the proprietary model GPT-4.1. SpatialReward's scores come from proprietary models, and an open-weight model only draws the edit boxes and filters the written reasons. ADIEE is the computed extreme. Its scores come from ground-truth edits and embedding similarities, at large scale, and they carry the biases of those metrics [24]. EditScore replaces those formulas with one teacher, GPT-4.1, and shows that the resulting reward model can pick the best of several edits and can drive online reinforcement learning [4]. SpatialReward tells its teachers, from the GPT and Gemini families, to attend to where the edit should happen, and then trains on those scores [7]. The PVC training set mixes both patterns: automatic metrics plus a proprietary model, on preservation pairs [22]. But none of these four training sets contains a human judgment. A score computed inside the training set therefore cannot show agreement with human raters. That check needs a separate test with human reference labels. Whether teacher scores are better or worse supervision than human ratings is outside what this survey can answer.

C. Independent Meta-Evaluation

That separate human test is not the same thing on every row of Table I, because the inclusion rule admits two uses. Of the sixteen sets, nine train a judge, some of them together with a test set from the same project. The other seven do not train a judge, and are only used to evaluate one. Some of the nine bring their own test. An internal test is a held-out part of the same project's data, in the sense above. In the cases here, that held-out part also shares sources, prompts, tasks, or candidate models with the training set. For example, Wu-EditReward-Bench shares that pipeline with EditReward-Data [6]. EBench-18K and MIE-Bench test in the same domain as their training data. In the released EditHF-1M rank splits, the same group key appears in training and validation [5]. That is leakage between the two splits, and the paper does not report its share. Hence, a score on such a test shows only that the judge is consistent inside one pipeline. That is internal validity, not evidence that the judge transfers to a different domain [31]. Torralba and Efros documented this gap for vision datasets [32]. The seven sets that only evaluate a judge are a different case.

All seven of them have human labels. Five come without a judge training set from the same project. GenAI-Bench editing votes come from an open arena [25]. AURORA-Bench builds its own tasks [2]. The MMRB2 editing subset was labeled for that benchmark by professional annotators [9]. EditReward-Compass is annotated independently, and many of its pairs draw both candidates from one model, close to a reinforcement-learning comparison [21]. EditInspector labels edits from MagicBrush with crowd raters [26]. MagicBrush is an editor training set, not a judge training set [10]. The other two, MER-Bench and VCReward-Bench, each sit next to the

TABLE I. SIXTEEN SETS FOR TRAINING OR TESTING IMAGE-EDITING JUDGES. READ THE RAW-ANNOTATION COLUMN BEFORE THE DERIVED-LABEL COLUMN: COUNTS KEEP THEIR ORIGINAL UNIT (GROUPS, EDITS, RATINGS, MEAN OPINION SCORES, OR PAIRS) AND ARE NOT COMPARABLE ACROSS ROWS WITHOUT THE PIPELINE IN FIG. 1. PROVENANCE MEANS WHERE THE LABELS COME FROM. ABBREVIATIONS: DH = DIRECT HUMAN, AH = AGGREGATED HUMAN, PT = PROPRIETARY TEACHER (A MODEL OUTSIDE THE TRAINING SET WRITES THE SCORES; FOR EDITSCORE THIS IS GPT-4.1), HC = HUMAN JUDGMENTS OF A WRITTEN CRITIQUE. OTHER ABBREVIATIONS: IF = INSTRUCTION FOLLOWING, VQ = VISUAL QUALITY, RM = REWARD MODEL, RL = REINFORCEMENT LEARNING, QA = QUESTION ANSWERING, SFT = SUPERVISED FINE-TUNING, GRPO = GROUP RELATIVE POLICY OPTIMIZATION, ECDF = EMPIRICAL CUMULATIVE DISTRIBUTION FUNCTION. YEAR IS THE YEAR OF THE FIRST PUBLIC RELEASE. ACCESS AND LICENSE REFLECT AUGUST 2026. A CODE LICENSE NEVER EXTENDS TO THE DATA OR THE WEIGHTS.

Set	Year	Raw annotation unit	Aggregated / derived labels	Prov.	Use	Overlap and caveats	Access, data license
HIVE reward set [19]	2023	1 human ranking per group of 5 edits; 3,634 groups, 1,412 rankable	10 ordered pairs per ranking; 1,285 train / 127 val groups	DH	RM training	candidates from one early editor; single rater; no external meta-eval	no separate release found; no license found
GenAI-Bench editing [25]	2024	1,154 anonymous arena votes	919 released pairs; four classes (A, B, tie-good, tie-bad)	DH	external meta-eval	arena filtering; rater expertise unknown	released; MIT in the paper, CC BY 4.0 on the card
AURORA-Bench [2]	2024	human scores (0/50/100) and pairwise preferences; 400 cases, 8 tasks	aggregation rule not documented; Fleiss κ 0.626 reported	DH	external meta-eval	small, curated; task-wise closeness to training sources varies	released; MIT (third-party sources separate)
EditInspector [26]	2025	3 crowd labels per edit on 783 MagicBrush edits	majority labels per dimension	AH	diagnostic bench	MagicBrush-only; mask-guided editors	released; CC BY 4.0 in the paper
EditReward-Data [6]	2025	1 expert rating per edit (IF, VQ on 1–4); 9,557 source-instruction cases	about 200K score-difference pairs (paper); 170,772 released rows	DH	RM training	same source and candidate pipeline as Wu-EditReward-Bench	released; data CC BY-NC 4.0, code MIT
EditScore training set [4]	2025	GPT-4.1 scores and rationales; about 70K samples	none	PT	RM training, best-of-N, RL	teacher bias; curated input and candidate selection	released; Apache-2.0
EBench-18K [20]	2025	1,101,600 judgments; 30 annotators; 15 per image and dimension	55,080 MOS + 18,360 QA labels over 18,360 edits	AH	judge training + internal bench	4:1 split unit unclear; in-domain evaluation	accessible; no data license stated
MMRB2 editing subset [9]	2025	3 professional raters; 7-point preference per pair; 1,000 pairs	majority label; ties and high-variance pairs removed	AH	independent meta-eval	near-unanimous model-judge pairs dropped	released; CC BY-NC 4.0, benchmark-only
EditHF-1M rank split [5]	2026	3 rankers per group (+3 on low agreement); 44.3K groups, 1.01M edits	29.1M dimension pairs from averaged ranks	AH	RM training + internal bench	train and validation not group-disjoint; tie rule undocumented	accessible; code MIT, no own data license
EditHF-1M score split [5]	2026	rater scores; 5 raters regular, 15 in test split	148K aggregated dimension scores (0–100)	AH	RM training + internal bench	split unit unclear	as above
EditReward-Compass [21]	2026	direct pair verdicts; 5 verifiers	2,251 pairs kept (5/5 unanimous); pre-filter pool not reported	DH	external meta-eval	ambiguity removed by unanimity filter; mostly intra-model pairs; some cross-model	accessible; no license stated
SpatialReward training set [7]	2026	teacher scores with spatial priors (Qwen/GPT/Gemini pipeline)	260K SFT + 7K GRPO samples	PT	RM training	teacher bias; multi-edit task focus	released; Apache-2.0
MER-Bench [7]	2026	5 trained experts; tiered rankings	600 groups / 1,800 images; 200 groups each with 2, 3, 4 edits	AH	external meta-eval	close in task type to a 60K multi-edit training set	accessible; no data license stated
VCReward-Bench [22]	2026	expert pair verdicts; strict preferences kept	3,506 released pairs (appendix sums to 3,508)	DH	preservation meta-eval	21 task types; bounded candidate pool	released; no license stated
ReasonEdit reward set [23]	2026	1,366,776 human judgments on 113,898 model critiques	ECDF/probit-normalized critique scores	HC	rationale-RM training	model-written critique pool; four accepted expert raters	data not released; code MIT
MIE-Bench [27]	2026	1.62M ratings; 15 raters per edit and dimension	108K MOS over 36,000 edits (3,000 cases $\times$ 12 editors)	AH	judge training + internal bench	3:1:1 split; new multi-source domain; no independent replication	accessible; no license stated

judge training set from the same paper [7], [22]. They are external when the judge under test was trained on other data. For the reward model released in the same paper, however, they are companion tests, so a score on them is not evidence of transfer.

So an internal test cannot show transfer. A controlled study would train judges on one set and test them on another, at comparable scale and with no shared sources. We found no such study for editing judges. The nearest substitutes are cross-benchmark tables in EditHF-1M and MIEScore. They test judges trained on one resource against the benchmarks of the others and leave out each judge's own training cell. Yet they mix model families and share image sources, so they do not isolate transfer. For reward models in vision-language understanding, a neighboring task, that controlled design exists. Average accuracy falls from 91.4 when training and test data come from the same preference dataset to 68.1 when they come from different ones, and one training set falls below chance [33]. That drop is evidence about those models rather than a measurement of editing judges. For editing judges, that measurement is missing.

D. Difficulty Shifts and New Label Objects

The usual question in the sets above is which edit is better overall. Several 2026 benchmarks change the comparison around that question, in three ways. The first way changes

the candidates. EditReward-Compass draws both candidates from one model for many of its pairs, while the rest compare two models [21]. The second way changes the editing task. MER-Bench uses one instruction that asks for several edits, and counts a group only when the full ranking is right [7]. VCReward-Bench asks only which edit preserves untouched content better, with instruction following held under control [22]. MIE-Bench combines two to five source images and rates attribute preservation in one score per edit [27]. The third way changes what is labeled. ReasonEdit scores the judge's written explanation, not the edit [23]. Liu et al. [30] rate one hundred HumanEdit cases on twelve factors, covering preservation, edit quality, and instruction fidelity, with twenty-five annotators. In short, these benchmarks are not substitutes for one another, and they are not substitutes for a single overall-preference score.

E. Access and Licensing

So far this section has asked what a label supports and what a test score means. Whether someone else can use the data is a different problem. In practice, a paper's claim that its data are open is not a reliable description of the release. Across the resources we checked, access falls into four states.

1) Released under a permissive license: Wu-EditReward-Bench (Apache-2.0), the EditScore training set and Luo-EditScore-Bench (Apache-2.0), the SpatialReward training set (Apache-2.0), AURORA-Bench (MIT), GenAI-Bench (MIT in the paper, CC BY 4.0 on the dataset card), and EditInspector (CC BY 4.0 in the paper).
2) Released with a non-commercial restriction: EditReward-Data (CC BY-NC 4.0) and the MMRB2 editing subset (CC BY-NC 4.0, benchmark-only use).
3) Reachable, but with no data license we could find or verify: EBench-18K, EditHF-1M, EditReward-Compass, MER-Bench, VCReward-Bench, and MIE-Bench. Counted as rows of Table I, that is seven of sixteen sets. EditHF-1M contributes two rows, rank and score.
4) Core data not published: the HIVE reward rankings, the ADIEE training set, the PVC training set, and the ReasonEdit human judgments.

TIEdit fits none of these four states. An archive appeared on the Hugging Face account that also hosts MIE-Bench on 12 August 2026, without a dataset card or license, and the paper links no release [28]. So two consequences follow from that list. A repository is not a data release, and a code license such as MIT or Apache-2.0 says nothing about the data or the weights beside it. The phrase "no license found" records our check at one time. It does not prove that no license exists.

V. Why Evaluation Numbers Do Not Travel

A license says whether the data may be used, but it does not say what an accuracy number measures. Judges in this area are routinely compared by a number called accuracy or correlation. Table II, on the next page, shows why that number does not travel from one benchmark to another. The protocols differ in what is compared, how many alternatives there are, what a tie means, how the human reference labels are built, which metric is reported, and whether the scored items depend on each other. The table uses the names from the introduction.

A. Comparison Units

Table II lines up those protocols. The thing being compared is not the same object on every row, so a single accuracy cannot rank them. The object runs from a score on one image (ImagenHub, EBench-18K), through a direct pair (MMRB2; EditReward-Compass also keeps a tie class) and a pair expanded from a ranking (Luo-EditScore-Bench, EditHF-1M), to a full group of candidates (Wu-EditReward-Bench, MER-Bench) and a separate label per error type (EditInspector). A judge that orders two candidates can still fail at a four-way ranking. Wu's strict metric counts a group as correct only when every pair inside it is correct, so the chance score falls as the group grows. A binary pairwise benchmark, by contrast, keeps chance at 50 percent. So putting those numbers on one leaderboard would compare quantities with different scales and different chance floors. Raji et al. make the related point that a benchmark measures the test it actually runs [35].

B. Tie Semantics

The comparison above is one reason a single accuracy cannot rank these benchmarks. The tie rule is a second reason. Benchmarks that keep ties score something closer to the full set of human labels, and benchmarks that drop ties score the clear preferences. The two families of numbers are not comparable, and a filter that removes ambiguity also removes the cases on which raters disagree most. GenAI-Bench keeps ties and separates a good tie from a bad tie [25]. EditReward-Compass keeps a unanimous tie and removes every other disagreement [21]. VCReward-Bench collects four options but keeps only pairs with a strict visual-consistency preference [22]. Luo-EditScore-Bench discards ties inside a tier [4]. MMRB2 removes ties and pairs on which the raters spread widely [9]. EditHF-1M does not say what happens when averaged ranks land on the same value [5]. Work on human label variation supports reading the disagreement as signal: much rater disagreement is systematic, and filtering it out discards part of what was labeled [36].

C. How the Reference Labels Are Built

These benchmarks do not report human agreement in the same way, so a judge's agreement with people on one of them is not on the same scale as its agreement on another. The human reference itself is built differently: one rater (HIVE, EditReward-Data), a majority of three raters or rater groups (EditInspector, MMRB2, Wu-EditReward-Bench), averaged rankings (EditHF-1M), a discussion between two experts until they agree (Luo-EditScore-Bench), agreement of all five verifiers (EditReward-Compass), and a mean of fifteen ratings (EBench-18K, MIE-Bench). The instruments differ as well. Rating scales and forced-choice comparisons do not have the same measurement variance, and they do not produce numbers on the same scale [37]. As a result, reported agreement differs

with the instrument. AURORA-Bench reports Fleiss κ of 0.626, a standard measure of how far a group of raters agrees beyond chance [2]. AbstractEdit, a 2026 benchmark for abstract editing intents, reports a moderate quadratic-weighted Fleiss κ of 0.47 [38]. Wu et al. report agreement per rating dimension [6]. Luo et al. compare their two-expert discussion with independent labeling and report over 97 percent agreement with a larger expert majority, a figure that is not comparable to a kappa over independent raters [4].

TABLE II. WHY A SHARED METRIC NAME DOES NOT MAKE TWO SCORES COMPARABLE. ROWS ARE GROUPED BY PROTOCOL FAMILY. CHANCE LEVELS ARE STATED ONLY WHERE THEY FOLLOW FROM THE NUMBER OF CLASSES. CLASSES / K LISTS THE ANSWER OPTIONS FOR ONE JUDGMENT: THE LABEL CLASSES OR THE RATING SCALE, AND, FOR A GROUP, THE NUMBER OF CANDIDATES K RANKED TOGETHER. DEPENDENCY: D = THE JUDGMENT WAS ASKED DIRECTLY, G = PAIRS WERE EXPANDED FROM ONE GROUP OR RANKING, R = THE SAME ITEM WAS RATED REPEATEDLY. D DESCRIBES HOW THE JUDGMENT WAS COLLECTED. IT DOES NOT CLAIM THAT THE JUDGMENTS ARE STATISTICALLY INDEPENDENT. SC/PQ = SEMANTIC CONSISTENCY AND PERCEPTUAL QUALITY. SRCC/PLCC = SPEARMAN AND PEARSON CORRELATION. PROTOCOL FACTS FOLLOW THE CITED PAPERS AND THE RELEASE CHECK (SECTION II). OVERLAP CAVEATS ARE IN TABLE I. A TIER IS A RANK LEVEL THAT SEVERAL CANDIDATES MAY SHARE.

Benchmark	Comparison unit	Classes / K	Tie handling	Reference construction	Metric (chance)	Dep.
Pointwise ratings, opinion scores, and mixed protocols						
ImagenHub [34]	pointwise rating per output	SC and PQ, each 0 / 0.5 / 1	n/a (pointwise)	3 raters; geometric mean of SC and PQ	rating correlation	R
AURORA-Bench [2]	pointwise score + pairwise preference	0/50/100	n/a	human raters; Fleiss κ 0.626 reported	correlation + pairwise agreement	D
EBench-18K [20]	pointwise MOS	0–100 continuous	n/a	15 ratings per image and dimension, normalized	SRCC / PLCC	R
Direct pairwise comparisons						
GenAI-Bench editing [25]	pairwise vote	4 (A, B, tie-good, tie-bad)	ties kept as two separate classes	anonymous arena votes, filtered	exact match (25 % uniform chance)	D
AugCLIP 2AFC [1]	two-alternative forced choice	2	no tie option	105 human raters	metric-human agreement (50 %)	D
AugCLIP triplets [1]	constructed triplet	2 per triplet	no tie option	contrasts from benchmark ground truth, source image, and generated over-edits	both-directions accuracy	n/a
MMRB2 editing [9]	pairwise preference	7-point scale $\rightarrow$ binary	ties and high-variance pairs (spread > 4) removed	3 professional raters, majority	pairwise accuracy (50 %)	D
EditReward-Compass [21]	direct pair	3 in the release (A, B, tie)	non-unanimous pairs removed; unanimous ties kept	5/5 unanimous verifier agreement	per-dimension accuracy	D
VCReward-Bench [22]	pairwise preference	2 (four options collected, strict preferences kept)	ties removed by the filter	expert verdicts	pairwise accuracy (50 %)	D
Pairs expanded from scores or rankings						
EditReward-Data pairs [6]	pair derived from score difference	2	tied overall scores split into two opposing training samples using the per-dimension scores	1 expert rating per edit	training signal, not a metric	G
Luo-EditScore-Bench [4]	pair from 5-way tier ranking	2	within-tier ties discarded	2 experts, discussed consensus	pairwise accuracy (50 % on kept pairs)	G
EditHF-1M rank [5]	pair expanded from group ranking	2	tie rule undocumented	averaged ranks of 3(+3) rankers	per-dimension pairwise accuracy (50 %)	G
Multi-way groups						
Wu-EditReward-Bench [6]	multi-way group	$K \in \{2, 3, 4\}$	hard cases kept deliberately	3 expert groups, majority	strict all-pairs-correct accuracy (chance falls with K)	G
MER-Bench [7]	multi-way tiered group	$K \in \{2, 3, 4\}$	overall-tier ties broken by the sub-scores; the reference is a strict order	5 experts, majority + review	multi-way accuracy + Kendall τ	G
Per-dimension error labels						
EditInspector [26]	per-dimension error label	varies by dimension	n/a	3 crowd labels, majority	accuracy / recall per dimension	R

TABLE III. WHICH FAILURE EACH BENCHMARK LABELS WITH HUMAN REFERENCE DATA. FILLED CIRCLES: HUMAN REFERENCE LABELS OR A FILTER. HALF CIRCLES: ONLY INSIDE A BROADER SCORE. OPEN CIRCLES: NO EXPLICIT LABEL. THE ASSIGNMENTS FOLLOW THE PUBLISHED PROTOCOLS. THEY IMPLY NO DIFFICULTY ORDER AND NO SHARED SCALE.

	instruction compliance	source preservation	modification balance	local artifacts	multi-source consistency	rationale quality
ImagenHub / VIEScore [3], [34]	◐	◐	○	◐	○	○
AURORA-Bench [2]	●	○	○	○	○	○
AugCLIP [1]	◐	◐	◐	○	○	○
EditInspector [26]	●	●	○	●	○	◐
Wu-EditReward-Bench [6]	●	◐	○	◐	○	○
Luo-EditScore-Bench [4]	●	●	○	◐	○	○
EBench-18K [20]	●	●	○	◐	○	○
MMRB2 editing [9]	◐	◐	○	○	○	○
EditReward-Compass [21]	●	●	○	◐	○	○
MER-Bench [7]	◐	◐	○	○	○	○
VCReward-Bench [22]	◐	●	◐	○	○	○
ReasonEdit [23]	◐	◐	○	○	○	●
MIE-Bench [27]	●	●	○	◐	●	○

D. Dependency Structure

Derived pairs from one ranking are not independent observations, which is why a pair count overstates how much information a dataset holds. EditReward-Data, EditHF-1M, and Luo-EditScore-Bench all expand pairs this way, marked G in Table II. If one averaged rank in a group of K candidates is wrong, that error enters up to $K-1$ derived pairs. In a group of five, for example, one wrong rank can touch four pairs. An uncertainty estimate that treats those pairs as independent therefore ignores the shared rank and can misstate how precise the estimate is. The ranked group, not the pair, is then the unit of analysis, and a bootstrap that resamples whole groups keeps the shared rank in the estimate. Under the Plackett–Luce model, Zhu et al. [39] show that splitting a K-way comparison into pairs still converges, but needs more data for the same precision than modeling the ranking directly.

VI. FAILURE MODES EXPOSED BY CURRENT BENCHMARKS

Each diagnostic benchmark isolates one way an editing judge can fail. Table III shows which benchmark attaches human reference labels to which failure, and which does not. The table records coverage. A filled circle means the benchmark attaches human reference labels, or a filter, to that failure. It does not mean the failure is harder there, and the symbols are not a quality ranking. An open circle means that failure is not labeled in that benchmark. The failure can still occur. The paragraphs below take one published finding at a time.

Editing is harder to judge than generation. On the ImagenHub labels, VIEScore measured rubric-prompted vision-language models against human reference labels, which later scorers also reuse, across generation and editing tasks. For GPT-4V, Gemini Pro, and LLaVA, correlation with those ratings was markedly lower on text-guided editing than on text-to-image generation. GPT-4o was the exception, while the open-weight LLaVA backbone correlates near zero on editing and stays well below human agreement even on text-to-image generation [3].

Similarity does not measure edit success. On AURORA-Bench, image similarity and human edit success come apart. AURORA-Bench scores edits with human reference labels for edit success (0, 50, or 100 for no, partial, and full execution) on action- and reasoning-centric tasks. Measured against these labels, pixel and embedding similarity mainly reward copying the source: a plain copy of the source image outscores every tested editor on these metrics, and CLIP-I correlates only 0.098 with human judgments on MagicBrush test edits [2].

Metrics can reward over-editing. An automatic similarity score can prefer the candidate that changes more than the instruction asked. AugCLIP builds contrast sets in which one candidate alters the source more than instructed. Directional CLIP similarity preferred the over-edited candidate in 63 percent of the TEdBench-based triplets and 35.1 percent of the MagicBrush-based triplets [1]. These experiments test similarity metrics rather than trained judges, but supervision pipelines that compute labels from such metrics (Section III) can carry this preference into training data; whether trained judges actually inherit it has not been measured directly in the surveyed corpus.

Artifact blindness. Here the judge misses a visible defect. EditInspector collects three crowd labels per edit and keeps the majority on each dimension (accuracy, artifacts, technical precision, consistency, difference captions), for 783 MagicBrush edits. In its artifact-detection test, GPT-4o reaches 65.7 percent accuracy but only 52.7 percent recall [26]: within this benchmark's domain, the model misses roughly half of the human-annotated artifacts. This holds for mask-guided MagicBrush edits, but it does not show that such models cannot see artifacts in general. Two neighboring results bound the finding. Encoder-level analyses show that CLIP-style vision backbones miss some fine visual differences [40], suggesting that encoder limitations can contribute to similar failures in CLIP-based

judges. In the text-to-image setting, human-annotated artifact regions can be learned by a trained predictor [41]. This establishes that artifact localization is learnable there, but does not identify the cause of the EditInspector result.

Source neglect. Here the judge scores the edit without looking at the changed region of the source image. SpatialReward uses that observation to motivate its design, and introduces MER-Bench, whose instructions ask for several edits and require a check at the level of the edited region [7]. Table III has no column for this failure, because the report comes from the judge's attention and not from human reference labels.

Preservation as a separate axis. VCReward-Bench isolates the preservation question: given two edits that both satisfy the instruction to a controlled degree, which one preserves the untouched content better [22]? Overall-preference benchmarks fold this question into a single score; VCReward-Bench makes it separately measurable.

Multi-source consistency. Here the edit must keep the right attributes from each of several source images. MIE-Bench extends the preservation question to edits that combine two to five source images, and rates attribute preservation in one score per edit [27]. No other surveyed benchmark attaches a dedicated human preservation score to multi-source edits.

Rationale quality. Here the object is the judge's written reason, not the edited image. ReasonEdit collects human reference labels for whether a critique is logical, accurate, and useful [23]. Those ratings measure explanation quality as the annotators see it, not whether the critique matches the model's actual decision process. A benchmark that only checks the score cannot detect an unfounded explanation. ReasonEdit is the only surveyed resource in which people rate the written explanation itself. EditInspector comes closest, since it compares model-written difference captions with human-written ones. For language models, chain-of-thought text can be plausible and still unfaithful to the model's decision [42]. That result is a neighboring finding. It is not a protocol that ReasonEdit runs.

Every finding above is bound to the domain, candidate construction, and human-reference protocol of its benchmark. Table III shows that these benchmarks label different failures and use incompatible protocols. Instruction compliance and source preservation appear in many rows. Multi-source consistency and rationale quality, by contrast, are each fully labeled by one benchmark. The twelve-factor benchmark of Liu et al. [30] moves toward joint coverage by rating preservation, edit quality, and instruction fidelity on one candidate pool. It covers the outputs of a single editing model and validates no rationales, so no surveyed benchmark measures all of these dimensions on a shared scale (Section VIII).

VII. Matching Data to Development Goals

The preceding sections separate label pipelines and evaluation protocols. This section turns that separation into a choice. Table IV matches a development goal to the resource types that fit it, and it names the caveat that belongs with each choice. The rows are alternatives for different goals. They are not a ranking. Scale and license of the listed sets are in Table I, and the failure coverage of the benchmarks is in Table III.

TABLE IV. Which data fit which development goal. Pick a row by the goal, and keep the caveat with any use of that data. The caveats follow from Sections IV to VI. The table does not measure which resource leads to a better judge.

Goal	Suitable resource types	Caveat
human-rated supervision at scale	EditReward-Data [6], EditHF-1M [5], EBench-18K [20]	separate raw ratings from derived pairs; check source/prompt/task overlap
scalable teacher supervision	EditScore [4], SpatialReward [7]	evaluate on a separate human reference set
heuristic training without a proprietary teacher	ADIEE-style pipelines [24]	metric bias must be stated; core data unreleased
independent overall preference	MMRB2 [9], GenAI-Bench [25]	sampling and tie filters must be reported
RL-near candidate pairs	EditReward-Compass [21]	unanimity filter removes ambiguity; many pairs are intra-model
artifact and preservation diagnosis	EditInspector [26], VCReward-Bench [22]	these benches are not interchangeable
multi-source editing	MIE-Bench [27]	internal train/test context; no independent replication
explaining judges	ReasonEdit [23]	critique quality is not image-judgment quality; core data unreleased

Three cross-cutting rules follow from the review. First, license compatibility is a selection criterion of its own. A team building a commercial judge should not assume that CC BY-NC data are compatible with commercial training and should obtain legal review. The risk is concrete: six of the seven accessible 2026 sets in Table I carry no stated data license. A large-scale study of fine-tuning corpora found the same pattern [43]. Second, whoever trains on a resource with an internal benchmark should report at least one number on an independent benchmark where one exists, because the internal number cannot show transfer. Third, whoever compares against published judge scores must quote the protocol with the number: a 55 percent pairwise accuracy on tie-filtered pairs and a 55 percent exact match over four classes describe different performance.

VIII. Open Problems and Recommendations

The comparisons above leave eight gaps. Five are missing facts in a release, and three are missing studies: no controlled transfer test, no benchmark that labels the failures together, and no joint check of score and explanation. Table V shows which of the nine reporting items each set states.

1. Preference-label pipelines lack a reporting standard for how rankings become pairs. Datasheets for Datasets ask how data were collected and labeled [44], but they do not ask for the rule that turns rankings or scores into pairs, or for how many independent judgments stand behind a derived pair.

TABLE V. THE NINE REPORTING ITEMS FOR THE SIXTEEN SETS. FILLED: STATED. HALF: ONLY IN PART. OPEN: NOT STATED. DASH: DOES NOT APPLY. PROVENANCE AND USE ARE LABELS GIVEN IN THIS SURVEY. DATA LICENSE: A LICENSE FOR THE DATA WAS STATED.

	raw count	aggregation	pair rule	provenance	protocol and ties	agreement	split and overlap	use	data license
HIVE	●	●	●	●	◐	○	◐	●	○
GenAI-Bench	●	◐	●	●	●	○	◐	●	●
AURORA-Bench	◐	◐	○	●	◐	●	◐	●	●
EditInspector	●	●	–	●	◐	●	◐	●	●
EditReward-Data	◐	●	●	●	●	○	◐	●	●
EditScore training	◐	◐	–	●	○	–	◐	●	●
EBench-18K	●	●	–	●	◐	●	◐	●	○
MMRB2 editing	●	●	◐	●	●	●	◐	●	●
EditHF-1M rank	◐	●	●	●	◐	●	◐	●	○
EditHF-1M score	○	◐	–	●	○	○	○	●	○
EditReward-Compass	◐	●	●	●	◐	●	◐	●	○
SpatialReward training	◐	◐	–	●	○	–	○	●	●
MER-Bench	◐	●	◐	●	◐	●	◐	●	○
VCReward-Bench	◐	◐	●	●	●	○	◐	●	○
ReasonEdit	●	●	–	●	○	○	○	●	○
MIE-Bench	●	●	–	●	◐	●	◐	●	○

2. The name "accuracy" covers incompatible measurements. Section V and Table II show the differences. A comparison across papers therefore needs the protocol restated with the number. RewardBench 2 gives text reward models an explicit tie test [45], but no surveyed editing benchmark tests whether a judge keeps equally acceptable edits close in score.

3. Cross-dataset transfer is not isolated for editing judges. A controlled train-on-A, test-on-B matrix was not found for image-editing judges (Section IV). For multimodal reward models in a neighboring task, that design shows an average drop of 23 accuracy points [33]. That drop, however, is not a measurement of editing judges.

4. Overlap checks stop too early. Disjoint sample identifiers do not guarantee disjoint sources, prompts, tasks, or candidate models, so a held-out split can still share the material the test was meant to leave behind. For example, the EditHF-1M rank splits are not group-disjoint between training and validation [5]. Hence, the four overlap axes should be checked separately.

5. Filters and overall scores hide information. Removing ties, non-unanimous pairs, or unrankable groups shifts a set toward clear preferences (Section V). So benchmarks should report how much data each filter removed, and report each dimension before any overall number.

6. Failure-mode labels are fragmented. The failures in Table III are labeled by separate benchmarks, so the checks cannot be put on one scale. In particular, no surveyed benchmark labels instruction compliance, preservation, artifacts, and explanations together.

7. Score and explanation are validated separately. No surveyed protocol checks the score and the explanation against human reference labels together. Yet ProJudgeBench shows that such labels can be built for multimodal reasoning [46].

8. Releases lack core data or a data license. Two of the sixteen sets in Table I have unreleased core data, and seven more are reachable but carry no stated data license. A code license does not cover the data or the weights [43].

The checklist below is the reporting rule for the five gaps that are missing facts.

Minimal reporting checklist. For each released set:

(i) raw judgment count and unit;
(ii) aggregation rule and rater population;
(iii) derived pair count and expansion rule;
(iv) provenance class and any mixture according to Fig. 1;
(v) comparison unit, K, tie rule, and chance level;
(vi) agreement statistic or consensus procedure;
(vii) the unit of the train/test split, plus overlap of sources, prompts, tasks, and candidate models;
(viii) intended use (training, internal benchmark, external meta-evaluation);
(ix) separate licenses for data, code, and weights.

IX. LIMITATIONS

This survey has six boundaries. First, the corpus is English-language arXiv and open-access work through August 2026. Work outside that record is not part of the claim, and the survey does not claim to be exhaustive. Second, many of these sets were released in 2026, during the writing of this survey. After the August 2026 check, a new file or license can change whether the data may be used. It does not change the score definition in the cited paper. The descriptions here are the versions we read, and this paper is not maintained as a living survey. Third, we did not re-run any judge on any benchmark. Every protocol and performance statement is taken from the cited papers and releases. Fourth, the access and license findings are a snapshot. The phrase "no license found" records the check in August 2026, and a release can change after that. Fifth, one author screened the corpus and coded the cells of Tables I, III and V. Neither step had a second pass, and the survey reports no record counts and releases no screening log. Sixth, the survey does not itself measure whether a judge trained on one set still works on another. The transfer gap in Section VIII is documented here, not measured.

X. Conclusion

A pair count is not the number of independent human decisions behind it, and an accuracy number measures one quantity only when the comparison, the tie rule, and the chance score are named with it. A test on data from the same project shows only that the judge is consistent there, not that it still works on other data. Likewise, a code license does not cover the data or the weights. The checklist in Section VIII turns these facts into nine items a release can report.

Acknowledgment

This research is funded by dtec.bw – Digitalization and Technology Research Center of the Bundeswehr. dtec.bw is funded by the European Union – NextGenerationEU.

References

[1] Y. Kim *et al.*, "Preserve or modify? context-aware evaluation for balancing preservation and modification in text-guided image editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025.

[2] B. Krojer *et al.*, "Learning action and reasoning-centric image editing from videos and simulations," in *Adv. Neural Inf. Process. Syst. (NeurIPS), Datasets and Benchmarks Track*, 2024.

[3] M. Ku, D. Jiang, C. Wei, X. Yue, and W. Chen, "VIEScore: Towards explainable metrics for conditional image synthesis evaluation," in *Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2024.

[4] X. Luo *et al.*, "EditScore: Unlocking online RL for image editing via high-fidelity reward modeling," in *Int. Conf. Learn. Represent. (ICLR)*, 2026.

[5] Z. Xu *et al.*, "EditHF-1M: A million-scale rich human preference feedback for image editing," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2026.

[6] K. Wu, S. Jiang, M. Ku, P. Nie, M. Liu, and W. Chen, "EditReward: A human-aligned reward model for instruction-guided image editing," in *Int. Conf. Learn. Represent. (ICLR)*, 2026.

[7] Y. Long *et al.*, "SpatialReward: Bridging the perception gap in online RL for image editing via explicit spatial reasoning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2026.

[8] J. Xu *et al.*, "ImageReward: Learning and evaluating human preferences for text-to-image generation," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023.

[9] Y. Hu, R. Askari-Hemmat, M. Hall, E. Dinan, L. Zettlemoyer, and M. Ghazvininejad, "Multimodal RewardBench 2: Evaluating omni reward models for interleaved text and image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2026.

[10] K. Zhang, L. Mo, W. Chen, H. Sun, and Y. Su, "MagicBrush: A manually annotated dataset for instruction-guided image editing," in *Adv. Neural Inf. Process. Syst. (NeurIPS), Datasets and Benchmarks Track*, 2023.

[11] S. Sheynin *et al.*, "Emu edit: Precise image editing via recognition and generation tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

[12] H. Snyder, "Literature review as a research methodology: An overview and guidelines," *Journal of Business Research*, vol. 104, pp. 333–339, 2019.

[13] X. Zang *et al.*, "Instruction-based image editing: A survey on data, models, evaluation, and applications," *Vicinagearth*, vol. 3, no. 1, 2026.

[14] Y. Huang *et al.*, "Diffusion model-based image editing: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 6, pp. 4409–4437, 2025.

[15] T. T. Nguyen *et al.*, "Instruction-guided editing controls for images and multimedia: A survey in LLM era," *arXiv:2411.09955*, 2024.

[16] J. Wang, J. Hu, X. Ma, H. Ma, X. Wei, and E. Wu, "Image editing with diffusion models: A survey," *arXiv:2504.13226*, 2025.

[17] J. Zhong *et al.*, "A comprehensive survey of reward models: Taxonomy, applications, challenges, and future," *arXiv:2504.12328*, 2025.

[18] D. Chen *et al.*, "MLLM-as-a-judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark," in *Int. Conf. Mach. Learn. (ICML)*, 2024.

[19] S. Zhang *et al.*, "HIVE: Harnessing human feedback for instructional visual editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

[20] Z. Xu *et al.*, "LMM4Edit: Benchmarking and evaluating multimodal image editing with LMMs," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2025.

[21] X. Bai *et al.*, "Edit-compass & EditReward-Compass: A unified benchmark for image editing and reward modeling," *arXiv:2605.13062*, 2026.

[22] Z. Jiang *et al.*, "GEditBench v2: A human-aligned benchmark for general image editing," *arXiv:2603.28547*, 2026.

[23] H. Chen, Z. Xu, H. Duan, X. Zhang, X. Min, and G. Zhai, "ReasonEdit: Towards interpretable image editing evaluation via reinforcement learning," *arXiv:2605.07477*, 2026.

[24] S. X. Chen, Y. Wei, L. Zhou, and S. Kumar, "ADIEE: Automatic dataset creation and scorer for instruction-guided image editing evaluation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2025.

[25] D. Jiang *et al.*, "GenAI arena: An open evaluation platform for generative models," in *Adv. Neural Inf. Process. Syst. (NeurIPS), Datasets and Benchmarks Track*, 2024.

[26] R. Yosef, M. Yanuka, Y. Bitton, and D. Lischinski, "EditInspector: A benchmark for evaluation of text-guided image edits," in *Proc. 63rd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2025.

[27] Z. Xu *et al.*, "MIEScore: Human-aligned evaluation for multi-source image editing," *arXiv:2608.02059*, 2026.

[28] S. Gao, Z. Xu, K. Fu, H. Duan, X. Min, and J. Wang, "Evaluating image editing with LLMs: A comprehensive benchmark and intermediate-layer probing approach," *Displays*, vol. 94, 2026, Art. no. 103494.

[29] F. Li, C. Wang, L. Lei *et al.*, "HP-Edit: A human-preference post-training framework for image editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2026.

[30] R. Liu, H. Weingord, S. Mittal *et al.*, "Human-aligned MLLM judges for fine-grained image editing evaluation: A benchmark, framework, and analysis," *arXiv:2602.13028*, 2026.

[31] T. Liao, R. Taori, I. D. Raji, and L. Schmidt, "Are we learning yet? A meta review of evaluation failures across machine learning," in *Adv. Neural Inf. Process. Syst., Datasets and Benchmarks Track*, 2021.

[32] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2011.

[33] Z. Li *et al.*, "The devil is in the details: Tackling unimodal spurious correlations for generalizable multimodal reward models," in *Proc. 42nd Int. Conf. Mach. Learn. (ICML)*, 2025.

[34] M. Ku *et al.*, "ImagenHub: Standardizing the evaluation of conditional image generation models," in *Int. Conf. Learn. Represent. (ICLR)*, 2024.

[35] I. D. Raji, E. Denton, E. M. Bender, A. Hanna, and A. Paullada, "AI and the everything in the whole wide world benchmark," in *Adv. Neural Inf. Process. Syst., Datasets and Benchmarks Track*, 2021.

[36] A. N. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, and M. Poesio, "Learning from disagreement: A survey," *Journal of Artificial Intelligence Research*, vol. 72, pp. 1385–1470, 2021.

[37] R. K. Mantiuk, A. Tomaszewska, and R. Mantiuk, "Comparison of four subjective methods for image quality assessment," *Computer Graphics Forum*, vol. 31, no. 8, pp. 2478–2491, 2012.

[38] M. Ventura, R. Hirsch, Y. Bitton, R. Cohen, and R. Reichart, "Editor's choice: Evaluating abstract intent in image editing through atomic entity analysis," *arXiv:2605.14842*, 2026.

[39] B. Zhu, M. I. Jordan, and J. Jiao, "Principled reinforcement learning with human feedback from pairwise or K-wise comparisons," in *Proc. 40th Int. Conf. Mach. Learn. (ICML)*, 2023.

[40] S. Tong, Z. Liu, Y. Zhai, Y. Ma, Y. LeCun, and S. Xie, "Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

[41] Y. Liang, J. He, G. Li *et al.*, "Rich human feedback for text-to-image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

[42] M. Turpin, J. Michael, E. Perez, and S. R. Bowman, "Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023.

[43] S. Longpre, R. Mahari, A. Chen *et al.*, "A large-scale audit of dataset licensing and attribution in AI," *Nature Machine Intelligence*, vol. 6, pp. 975–987, 2024.

[44] T. Gebru *et al.*, "Datasheets for datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021.

[45] S. Malik *et al.*, "RewardBench 2: Advancing reward model evaluation," in *Int. Conf. Learn. Represent. (ICLR)*, 2026.

[46] J. Ai, P. Zhou, Z. Xu *et al.*, "ProJudge: A multi-modal multi-discipline benchmark and instruction-tuning dataset for MLLM-based process judges," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2025.