AXISRUBRIC: RUBRIC CONSTRUCTION VIA AXIS DISCOVERY FOR JUDGING MULTIMODAL GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Rubric-based rewards improve comparative preference prediction for multimodal generation by decomposing human preferences into explicitly checkable evaluation axes. Their accuracy therefore depends on how effectively these axes capture the factors that shape human preference. A largely overlooked challenge in discovering preference-relevant axes for reward modeling in multimodal generation is the *cross-modal observability gap*. That is, textual instructions and predefined rubrics may miss delivery-quality differences, such as rendering artifacts, audio clarity, and motion coherence, which are observable only in the generated multimodal outputs. Existing methods use instruction-derived, expert-authored, or output-aware rubrics, yet none fully closes the gap. Our preliminary study shows that even when the rubric author sees the generated outputs, it still focuses on what the instruction requests, and up to 32% of human-validated delivery-quality axes are not discovered. Moreover, the discovered axes are not equally predictive of human preference, nor are they applicable to every instance. These observations reveal two key failure modes: *axis omission*, where preference-relevant axes are missing from the rubric; *dilution*, where redundant, poorly predictive, or inapplicable axes weaken the influence of informative ones. To close the gap, we propose AxisRubric, a three-phase multi-agent framework for automatic rubric construction via axis discovery. To address omission, Phase 1 runs independent instruction-blind and instruction-informed observer agents that discover evaluation axes, which we consolidate into an evolving benchmark-level taxonomy. To address dilution, Phase 2 deactivates misleading axis categories and calibrates the weights of the remaining axes, and Phase 3 selects the axes applicable to each individual instance and rewrites them as instance-specific rubric items. Across twelve preference benchmarks spanning image, audio, and video generation, AxisRubric outperforms all seven evaluated baselines. It improves preference accuracy over the strongest automatic rubric construction baseline by 0.08 on average (from 0.66 to 0.74) and by up to 0.18 (from 0.70 to 0.88). Notably, these gains hold across 21 pairs of rubric-author and output-judge models.

1 INTRODUCTION

With the rapid advancement of multimodal generative models, aligning generated audio, images, and videos with human preferences has become increasingly important. A common approach to preference-based alignment is to have humans compare candidate outputs generated for the same instruction and indicate which output they prefer, and then train the model to favor outputs that better align with these human preferences (Xu et al., 2023; 2026; Liu et al., 2025). Among these approaches, rubric-based rewards have emerged as a particularly promising framework (Gunjal et al., 2026; Jia et al., 2026a; Tian et al., 2026). In this framework, the comparison is guided by a *rubric*, a set of evaluation questions called *rubric items*, and the judge answers each rubric item before deciding which output is better. This rubric-based reward provides fine-grained supervision, improves agreement with human judgments, and makes reward decisions more interpretable.

Despite these benefits, a rubric can only compare outputs with respect to the aspects explicitly captured by its items. We refer to these aspects as *evaluation axes*. In multimodal generation, evaluation axes fall into two types: *instruction-following* axes, which check whether the output contains what the instruction requests (e.g., scene content and requested activity in Fig. 1); *delivery-quality* axes,

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.

which check the quality of the generated output (e.g., anatomical plausibility and lighting consistency in Fig. 1). If an axis matters to humans but is missing from the rubric, the rubric-judge agent never considers it, and the resulting reward cannot capture that aspect of human preference, regardless of the judge model's capabilities. Therefore, *a key challenge is to develop a generalizable method for automatically constructing rubrics that faithfully capture human preferences across diverse multimodal generation tasks*.

Existing approaches mostly construct rubrics in three ways, depending on what the rubric author is given (Fig. 1). First, *instruction-derived methods* write rubric items from the generation instruction alone (Hu et al., 2023; Cho et al., 2024; Li et al., 2026b). They are effective for checking whether an output follows the instruction, but they cover only the axes that the instruction states. Our empirical analysis reveals that they miss about 98% of the delivery-quality axes that humans rely on (see Insight 1). Second, *expert-authored rubrics* specify the axes in advance for a particular image, video, or audio task (Huang et al., 2024; Liu et al., 2024; Xu et al., 2026; Tjandra et al., 2025). They cover the delivery-quality axes that experts anticipated, but extending them to unanticipated axes or new tasks requires additional expert design. Third, *output-aware methods* go one step further by allowing the rubric author to examine the generated outputs and refine the original generated rubric (Kao et al., 2026; Tian et al., 2026; Yu et al., 2026). By examining the generated outputs, these methods recover many delivery-quality axes that the instruction leaves unstated. However, seeing the outputs is not enough. The rubric author is still biased toward the aspects requested by the instruction, even when outputs are shown alongside the instruction, leaving up to 32% of human-validated delivery-quality axes undiscovered (see Insight 1).

These limitations point to the same underlying problem. As shown in Fig. 1, the axes omitted by existing approaches (red dashed boxes) are all delivery-quality axes, which the instruction never mentions. Human preferences often depend on delivery quality, which can only be observed in the generated audio, image, or video, whereas instructions and predefined rubrics are written in text. In other words, the evidence and the rubric live in different modalities, and what humans expect can go beyond what the textual instruction captures. We refer to this cross-modal mismatch as the *cross-modal observability gap*. None of the existing approaches fully closes this gap, and each leaves a substantial share of preference-relevant axes undiscovered, resulting in what we call *axis omission*.

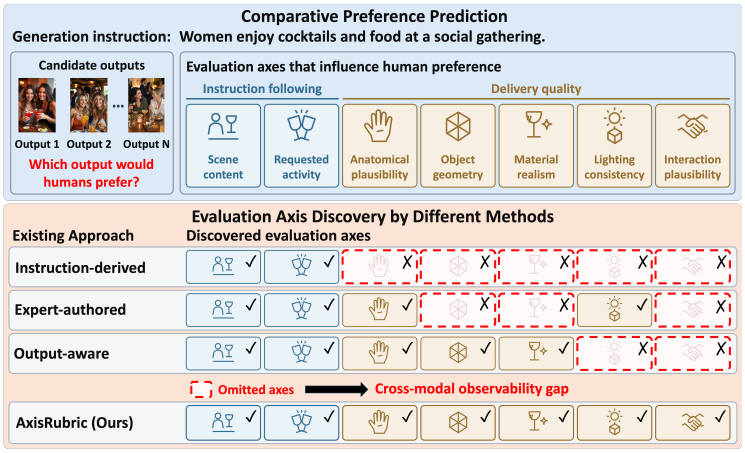

Figure 1: Illustration of the cross-modal observability gap.

Moreover, discovering the missing axes is not sufficient on its own. An overall human preference does not tell us how much each axis matters or which axes apply to a given instance. If all axes are treated equally, weakly informative or inapplicable rubric items receive the same weight as informative ones, weakening the final decision. We refer to this effect as *dilution*. Our empirical analysis reveals two sources of dilution (see Insight 2). Equal weighting lowers preference accuracy by up to 0.02 compared with calibrated weighting, and reusing one benchmark-level rubric for every instance lowers accuracy by up to 0.08 compared with per-instance rubrics. These observations highlight the need for a rubric construction method that discovers the missing evaluation axes and aligns their weights with human preferences. There are two key challenges to achieving this goal:

(**C1**) *How can we discover the missing evaluation axes that shape human preference from candidate multimodal outputs?* The discovery must preserve both delivery-quality and instruction-following information as complementary sources, without one filtering out the other. It must also consolidate recurring observations into distinct, minimally overlapping axes.

(**C2**) *How can we calibrate discovered evaluation axes to human preferences and generate rubrics on a per-instance basis?* The method must weigh each axis by how well it aligns with human preferences, while limiting the influence of misleading or weakly supported axes. It must then select the applicable axes and generate rubric items for each instance separately.

To address these challenges, we propose **AxisRubric**, a three-phase multi-agent framework that automatically constructs rubrics by discovering preference-relevant evaluation axes for comparative reward modeling in multimodal generation. We summarize our main contributions as follows.

First, to the best of our knowledge, this is the first systematic work that addresses the *cross-modal observability gap* in automatic rubric construction for reward modeling in multimodal generation. We identify two key failure modes, *axis omission* and *dilution*, and quantify their impact on preference accuracy across image, audio, and video generation.

Second, we propose AxisRubric to address (C1) and (C2), as illustrated in Fig. 2. The framework builds a *taxonomy*, a hierarchy with *categories* at the top, *subcategories* beneath them, and evaluation axes at the leaves. To address (C1), *Systematic Axis Discovery* (Phase 1) uses two observer agents, one instruction-blind and one instruction-informed, to record delivery-quality and instruction-following differences. It then consolidates recurring observations into distinct evaluation axes and updates the taxonomy as new observations are added. To address (C2), *Axis Weight Calibration* (Phase 2) uses human preferences to deactivate misleading or uninformative categories. It assigns axis-level weights based on preference accuracy and the amount of supporting evidence. *Per-Instance Rubric Generation* (Phase 3) then selects applicable axes and generates rubric items that specify what to reward or penalize under each generation instruction, enabling the rubric to reflect both human preferences and the requirements of each instance.

Finally, we evaluate AxisRubric on twelve preference benchmarks spanning image, audio, and video generation. AxisRubric outperforms all seven baselines and improves preference accuracy over the strongest automatic rubric-construction baseline in each modality (i.e., TIFA, AudioRubrics, and RAVEN-Eval) by 0.08 on average (from 0.66 to 0.74) and up to 0.18 (from 0.70 to 0.88). Notably, these gains hold consistently across 21 pairs of rubric-author and output-judge models.

2 Related Work

In this section, we first review reward models and evaluation methods for multimodal generation. We then organize rubric generation methods into three groups (instruction-derived, expert-authored, and output-aware rubrics) according to how their evaluation axes are derived.

Reward Modeling for Multimodal Generation. Reward models predict human preferences over generated outputs and provide the training signal for aligning multimodal generators with human preferences. Scalar reward models such as ImageReward (Xu et al., 2023), HPS (Wu et al., 2023), PickScore (Kirstain et al., 2023), VideoReward (Liu et al., 2025), and Omni-Reward (Jin et al., 2026) learn a single reward score from human annotations. To make the reward score interpretable, multimodal large language model (MLLM) judges, such as VIEScore (Ku et al., 2024) and T2I-Eval-R1 (Ma et al., 2025b), produce preference decisions with natural-language explanations. However, scores and explanations alone do not establish whether all preference-relevant quality axes have been considered. Rubric-based rewards, such as VisionReward (Xu et al., 2026) and RubricRL (Feng et al., 2026), make the evaluation axes explicit. They score each rubric item and aggregate the item scores into the reward, which reveals on which axes an output falls short. Their accuracy therefore depends on whether the rubric covers the axes that matter to human preferences. The next three paragraphs group existing methods by where their rubric axes come from.

Instruction-Derived Rubrics. The first direction derives rubrics primarily from the generation instruction. TIFA converts text-to-image instructions into questions answered by a visual question-answering model (Hu et al., 2023). DSG further organizes such questions into a dependency graph to reduce redundant checks and inconsistent answers (Cho et al., 2024). RubricRL generates a checklist for each instruction, including a few generic quality items such as realism, and aggregates item-level judgments into rewards (Feng et al., 2026). For audio, AnyAudio-Judge decomposes instructions into binary rubric items (Li et al., 2026b). For video, RAVEN-Eval appends a fixed set of basic quality criteria to the instruction-derived items (Jia et al., 2026b), and VideoArgus selects

criteria from a predefined set for each instruction and its reference inputs (Zeng et al., 2026). These rubrics contain only the requirements stated in the instruction and a few generic quality criteria fixed in advance. However, many delivery-quality axes that shape human preference are neither stated in the instruction nor included in these fixed criteria, so they fall outside these rubrics, resulting in axis omission (see Insight 1).

Expert-Authored Rubrics. To cover the quality requirements that instructions leave unstated, the second direction fixes the evaluation axes in advance based on expert knowledge of a modality. For video, VBench (Huang et al., 2024) and EvalCrafter (Liu et al., 2024) fix a set of video quality axes, such as subject consistency, motion smoothness, and text-video alignment. For image and video, VisionReward organizes a set of binary questions and learns weights to combine their answers (Xu et al., 2026). For audio, Audiobox Aesthetics predicts human ratings on four predefined quality axes (Tjandra et al., 2025). UEval constructs a rubric for each question from the reference inputs and has human experts validate it (Li et al., 2026a). While these approaches cover quality axes beyond the instruction, they capture only the axes that experts anticipated, and extending them to new failure modes or tasks requires more expert effort.

Output-Aware Rubrics. The third direction constructs rubrics from both instructions and generated outputs. AutoRubric-T2I (Kao et al., 2026) and ARR (Tian et al., 2026) derive rubric items from preference-labeled image pairs, where AutoRubric-T2I consolidates them into one rubric shared across instances and ARR generates a rubric for each instance. AudioRubrics generates a rubric for each question and updates it with on-policy generations during training (Yu et al., 2026). RubricRM (Kan et al., 2026) and DyCoRM (Qian et al., 2026) fine-tune a single reward model on annotated image pairs. The model generates instance-specific axes from each instruction and its output candidates, and then scores the candidates with the same model. The shared rubric of AutoRubric-T2I applies the same items to every instance and thus suffers from dilution (see Insight 2), which the other methods mitigate by adapting the rubric to each instance. Nevertheless, none of these methods closes the cross-modal observability gap, since the instruction still biases rubric generation toward its stated requirements even when the generated outputs are observed (see Insight 1).

3 KEY INSIGHTS

The cross-modal observability gap limits rubric coverage and affects preference accuracy in multimodal reward modeling. We present two key insights that motivate the design of AxisRubric.

Insight 1: Effective multimodal rubrics require systematic discovery of preference-relevant evaluation axes. In multimodal generation, human preferences depend on both instruction following (IF) and delivery quality (DQ). *Axis omission* occurs when a rubric misses evaluation axes relevant to human preferences. We examine axis omission across three preference benchmarks for each of image, audio, and video generation, using the same observer model and 40 discovery instances per benchmark. We compare (i) instruction-derived discovery using the generation instruction only, (ii) instruction-blind observation using the generated outputs only, and (iii) joint observation using both. For each benchmark, we pool observations from all three settings and consolidate similar ones into distinct IF and DQ axes. Three LLM judges then evaluate each axis on held-out instances, and axes whose decisions predict human preference labels better than chance form our *reference set*. As shown in Table 1, instruction-derived discovery recovers only 1.4-2.0% of reference DQ axes, as instructions may leave delivery expectations unstated. In contrast, observing the outputs alone raises DQ discovery to about 80%, but recovers at most 25.4% of reference IF axes. Providing both recovers the IF axes, but joint observation still recovers only 68.3-71.3% of reference DQ axes, about 9-11% fewer than instruction-blind observation. This is because including the instruction shifts discovery toward IF. Therefore, none of the three discovery settings alone recovers all reference axes, leaving the cross-modal observability gap unresolved. These findings motivate systematic discovery of both IF and DQ axes relevant to human preferences to close the cross-modal observability gap.

Insight 2: The discovered axes are neither equally predictive of human preference nor applicable to every instance. Recovering missing axes is not enough. When poorly predictive or inapplicable axes count as much as informative ones, they weaken the informative ones, which we call *dilution*. Dilution arises in two ways. First, some axes predict human preferences poorly, and

Table 1: Reference-set coverage of the three discovery settings, averaged over three benchmarks per modality (listed in Appendix A.1). The reference set columns give the average number of IF and DQ reference axes, and the remaining columns give the percentage recovered by each setting.

Modality	Reference set (#)		Instruction-derived (%)		Instruction-blind (%)		Joint observation (%)	
	IF	DQ	IF	DQ	IF	DQ	IF	DQ
Image	9.0	18.0	55.6	1.7	0.0	79.4	52.2	68.3
Audio	6.0	20.7	88.3	1.4	5.0	82.1	83.3	71.0
Video	6.7	15.0	74.6	2.0	25.4	80.0	94.0	71.3

Table 2: Average preference accuracy across three benchmarks per modality.

Weighted	Rubric generation	Image	Audio	Video
✓	Per-benchmark	0.68	0.66	0.66
✗	Per-instance	0.74	0.71	0.71
✓	Per-instance	**0.76**	**0.73**	**0.73**

equal weighting lets them sway the prediction as much as informative axes. Second, some axes are applicable to certain instances but not others, so a rubric shared across instances lets the inapplicable ones sway the prediction. To measure each effect, we start from per-instance rubrics with axis weights calibrated on human preferences and remove one component at a time: (i) removing weight calibration, so that all axes receive equal weight; (ii) removing per-instance generation of rubric items, so that one benchmark-level rubric is used for all instances. In each setting, we score every candidate output by the weighted sum of its rubric-item judgments, select the highest-scoring one as the preferred output, and count the prediction as correct if it matches the human label. We report average preference accuracy across three benchmarks per modality, which are listed in Appendix A.1. As shown in Table 2, removing weight calibration lowers average accuracy by up to 0.02, and removing per-instance generation by up to 0.08. These findings motivate us to mitigate dilution by aligning axis weights with human preferences and adapting the rubric items to each instance.

4 Our Design: AxisRubric

In this section, we present AxisRubric, an automatic three-phase framework for constructing rubrics for reward modeling in multimodal generation. Leveraging the aforementioned insights, we aim to discover the evaluation axes that humans rely on and to weigh them according to human preferences. As shown in Fig. 2, *Systematic Axis Discovery* (Phase 1; Section 4.1) discovers preference-relevant evaluation axes, addressing (C1) by reducing *axis omission* (Insight 1). Since the discovered axes differ in how well they capture human preferences, *Axis Weight Calibration* (Phase 2; Section 4.2) calibrates the weight for each axis. Since a benchmark-level rubric may still include axes that do not apply to a particular instance, *Per-Instance Rubric Generation* (Phase 3; Section 4.3) selects applicable axes and constructs rubric items specifying what to reward or penalize. Together, Phases 2 and 3 address (C2) by mitigating *dilution* (Insight 2). For each benchmark, we use a set of training instances for Phases 1 and 2 and a separate held-out set for evaluation.

4.1 Phase 1: Systematic Axis Discovery

Each training instance i contains a generation instruction x_i, $N_i \geq 2$ candidate outputs, and a human preference label indicating which candidate output humans prefer. We discover evaluation axes from the training instances by alternating between observation and taxonomy evolution over multiple rounds. In each observation round, two independent LLM observer agents identify differences among the candidate outputs of each instance. The *instruction-blind observer agent* receives only the candidate outputs and identifies DQ differences such as blurry images, background noise, or jerky video motion. The *instruction-informed observer agent* additionally receives x_i and identifies both IF and DQ differences relative to the instruction. Neither agent sees the human preference label or the other's observations. Each difference is recorded as an *observation* with a label, a description, a severity, and a confidence score. To organize observations into axes, we maintain a *taxonomy*,

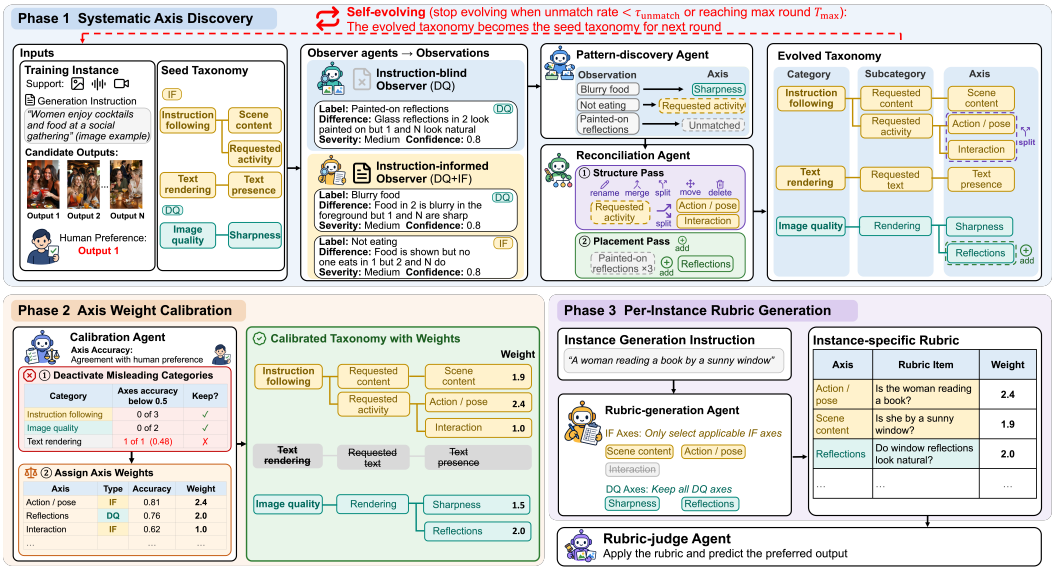

Figure 2: Overview of AxisRubric.

a three-level tree of categories, subcategories, and evaluation axes. The first round starts from a small *seed taxonomy*, which an LLM drafts from the evaluation aspects we specify, and each later round starts from the taxonomy evolved in the previous round. Each observer agent matches every observation to an existing axis whenever one fits, and places the rest in the *unmatched* category.

After each observation round, we use two LLM agents to evolve the taxonomy by refining existing axes and creating new axes for recurring unmatched observations. A *pattern-discovery agent* first checks the observations under each axis to find axes that are too broad, unclear, or overlapping with another. It then groups the unmatched observations by similar difference they describe, which reveals recurring differences that no axis covers. A *reconciliation agent* then updates the seed taxonomy of the round based on these grouped observations in two passes (Fig. 2). The *structure pass* revises existing axes with five operations: *rename* an unclear axis, *merge* overlapping axes, *split* an overly broad axis, *move* an axis to a better position in the taxonomy hierarchy, and *delete* an axis that is no longer needed. The *placement pass* only adds axes: for each group of unmatched observations that recurs at least twice, it adds a new axis with a name and a definition under the best-fitting subcategory. Each observation and axis carries an IF or DQ tag. The IF and DQ axes are never merged or placed under the same category, ensuring that these two types of axes remain separate for the applicability check in Phase 3. We repeat observation and taxonomy evolution for at most $T_{\max}$ rounds, or until the unmatch rate (i.e., the fraction of unmatched observations) falls below τ_{unmatch}. We provide the prompts of both agents in Appendix A.5 and set $T_{\max}$ and τ_{unmatch} via grid search on the training instances (see Appendix A.4).

4.2 Phase 2: Axis Weight Calibration

After taxonomy evolution, an LLM calibration agent evaluates each axis on the labeled training instances, one axis at a time. For each instance, the agent selects the best candidate output according to the axis definition, and repeats the selection under a second presentation order (i.e., the reversed order when $N_i = 2$). The prediction is correct only when both orders select the human-preferred candidate. A disagreement between the two orders counts as incorrect, as the axis provides no preference-relevant signal in this case. We then compute the *axis accuracy* $p_\ell \in [0, 1]$ of each axis ℓ as the fraction of training instances it predicts correctly, and the *category accuracy* p_c of each top-level category c as the fraction of correct predictions among all instance-axis pairs in c. With these accuracies, we first deactivate misleading categories and then weigh the axes, ignoring the unmatched category in both steps as it contains no evaluation axes.

Category deactivation. We deactivate a category and all of its axes when more than a fraction ρ of its axes have accuracy below an accuracy threshold τ, where ρ is selected via grid search on the training instances (Appendix A.4). We deactivate at the category level rather than at the axis level, since a single axis accuracy estimated from a few training instances is noisy, whereas low accuracy across many axes of a category is stronger evidence that the category is misleading. If no category would remain with category accuracy $p_c \geq \tau$, we retain all categories to preserve potentially useful combinations of axes. The resulting active categories form $\mathcal{C}_{\text{act}}$, and their axes form $\mathcal{L}_{\text{act}}$.

Axis weighting. We assign each active axis a weight w_ℓ in proportion to its axis accuracy p_ℓ and its category accuracy p_c. Since the two types of accuracy lie on different scales, we normalize each axis accuracy by the highest axis accuracy among all active axes and each category accuracy by the highest category accuracy among all active categories. Specifically, given an axis or category accuracy p and the corresponding highest accuracy $p_{\max}$, we compute a multiplier $m(p)$ as

$$m(p) = 1 + (M-1)\,\frac{\max(p-\tau, 0)}{p_{\max}-\tau}, \tag{1}$$

which equals 1 when p is no greater than the threshold τ, increases linearly with p above τ, and reaches its maximum M at $p_{\max}$. Here, M is the largest multiplier an axis or a category can receive, which is selected via grid search on the training instances (Appendix A.4). Then, each active axis ℓ receives an *axis multiplier* $m(p_\ell)$, and each active category c receives a *category multiplier* $m(p_c)$ shared by all axes in the category. The weight of axis ℓ is $w_\ell = m(p_c)\,m(p_\ell)$, i.e., the product of two multipliers. The calibrated axes and weights are fixed per benchmark. Phase 3 only selects the applicable axes for each instance and rewrites them as rubric items, without changing their weights.

4.3 Phase 3: Per-Instance Rubric Generation

For each evaluation instance i, an LLM rubric-author agent constructs a rubric from the instruction x_i and calibrated taxonomy, without seeing any generated output. The agent first decides which axes are applicable. It checks each active IF axis against x_i and excludes the inapplicable ones (e.g., a lyric-related axis for instrumental music), while keeping all active DQ axes, as delivery expectations affect human preferences even though the instruction may not explicitly specify them. Let $\mathcal{S}_i$ denote the retained axes. For each retained axis $\ell \in \mathcal{S}_i$, the agent then writes an instance-specific rubric item $q_{i,\ell}$, which specifies what to look for and what counts as an error under x_i. To do so, it keeps the meaning of the axis, but refers to objects, actions, styles, or sounds in the instruction. For example, distortion is an error for clean speech but a desired feature when the instruction asks for it. Each item inherits the calibrated weight w_ℓ of its axis, and the resulting rubric is $\mathcal{R}_i := \{(q_{i,\ell}, w_\ell) : \ell \in \mathcal{S}_i\}$. Finally, a rubric-judge agent, using an LLM different from the rubric author, applies $\mathcal{R}_i$ to predict the preferred output (see Appendix A.4 for the judging protocol).

5 Evaluation

We evaluate AxisRubric on twelve preference benchmarks across image, audio, and video generation. We use a pair of candidate outputs per instance since most benchmarks only provide pairs.

5.1 Experimental Setup

Datasets. We conduct experiments on six image benchmarks: Pick-a-Pic (Kirstain et al., 2023), HPS v2 (Wu et al., 2023), HPS v3 (Ma et al., 2025a), ImageRewardDB (Xu et al., 2023), MJ-Bench (Chen et al., 2024), and OmniReward-T2I (Jin et al., 2026), three audio benchmarks: Rapidata (Rapidata, 2025), Music-Arena (Kim et al., 2025), and OmniReward-T2A (Jin et al., 2026), and three video benchmarks: VideoGen-RewardBench (Liu et al., 2025), GenAI-Arena (Jiang et al., 2024), and OmniReward-T2V (Jin et al., 2026). We run AxisRubric separately on each benchmark, using the same 40 instances for both axis discovery and weight calibration and a held-out set for evaluation, ensuring that the taxonomy and weights are constructed and evaluated within the same benchmark. We provide details about the benchmarks and data-splitting procedure in Appendix A.1.

Baselines. We compare AxisRubric with seven baselines that differ in where their evaluation axes come from. *Holistic* asks the judge for an overall preference without a rubric. *Fixed* uses expert-authored rubrics from VisionReward for image and video (Xu et al., 2026) and those from Audiobox

Table 3: Preference accuracy with Gemini 2.5 Pro as the rubric author and Gemini 3.5 Flash as the judge. Bold marks the best result, and underline marks the second-best per benchmark.

Image Benchmark	Holistic	Fixed	Decomposition	TIFA	AutoRubric-T2I	AxisRubric
Pick-a-Pic	0.53	0.60	<u>0.70</u>	0.69	0.51	**0.73**
HPS v2	0.73	<u>0.78</u>	0.75	<u>0.78</u>	0.74	**0.80**
HPS v3	0.53	<u>0.70</u>	0.55	0.59	0.64	**0.73**
ImageRewardDB	0.68	<u>0.70</u>	0.65	0.69	0.68	**0.78**
MJ-Bench	0.60	0.58	0.65	<u>0.75</u>	0.61	**0.78**
OmniReward-T2I	0.60	0.60	0.50	0.58	<u>0.62</u>	**0.63**

Audio Benchmark	Holistic	Fixed	AnyAudio	AudioRubrics	AxisRubric
Rapidata	<u>0.57</u>	0.53	0.50	0.48	**0.58**
Music-Arena	<u>0.83</u>	0.78	0.60	0.70	**0.88**
OmniReward-T2A	0.58	0.55	0.63	<u>0.67</u>	**0.73**

Video Benchmark	Holistic	Fixed	Decomposition	TIFA	RAVEN	AxisRubric
VideoGen-RewardBench	0.63	<u>0.73</u>	0.55	0.50	0.69	**0.75**
GenAI-Arena	<u>0.68</u>	0.63	0.43	0.60	<u>0.68</u>	**0.70**
OmniReward-T2V	0.65	<u>0.70</u>	0.55	0.60	0.65	**0.75**

Aesthetics for audio (Tjandra et al., 2025). The remaining baselines construct rubrics automatically. *Decomposition* derives the rubric from the generation instruction, following AnyAudio-Judge (*AnyAudio*) for audio (Li et al., 2026b). We also include *TIFA* on image and video (Hu et al., 2023), *AutoRubric-T2I* on image (Kao et al., 2026), *AudioRubrics* on audio (Yu et al., 2026), and RAVEN-Eval (*RAVEN*) on video (Jia et al., 2026b). For a fair comparison, we keep each method's rubric-construction procedure, omit its model training and downstream pipeline, and evaluate the resulting rubric with the same judge model and judging protocol as AxisRubric (see Appendix A.2 for details). Finally, we include three benchmark-trained reward models for image: PickScore, HPSv2, and ImageReward, which rank the two candidate outputs by their scalar scores without a judge.

Metrics. We report preference accuracy, the fraction of preference predictions that match the human preference labels. To examine robustness, we further assess judging consistency by reporting the self-consistency rate and Fleiss' κ of the judge model across three repeated judgments with the same rubrics (see Appendix D.1), and measure the change in accuracy when a fraction of the calibration labels is flipped (see Appendix D.2). We provide full details of these metrics in Appendix A.3.

Implementation Details. We use Gemini 2.5 Pro as the rubric author and Gemini 3.5 Flash as the judge in the primary configuration and report the results across all twelve benchmarks in Table 3. To test whether our gains depend on particular models, we also use Gemini 2.5 Pro, Gemini 3.5 Flash, and Qwen3-Omni-30B-Thinking in both roles, and Claude Sonnet 5 as an additional rubric author for image. The rubric author and judge are always different models, yielding nine author-judge configurations for image and six each for audio and video. We provide the hyperparameters, experimental settings, and judging protocol in Appendix A.4 and the agent prompts in Appendix A.5.

5.2 Main Results

Consistent Gains Across Multimodal Benchmarks. As shown in Table 3, AxisRubric achieves the highest accuracy on all twelve benchmarks, with an overall accuracy of 0.74. To quantify the improvement, we compare AxisRubric with the strongest automatic rubric baseline in each modality by average accuracy, namely TIFA for image, AudioRubrics for audio, and RAVEN for video. AxisRubric improves over these baselines by 0.06, 0.11, and 0.06 on average, and by 0.08 (from 0.66 to 0.74) when averaged over the twelve benchmarks, and by up to 0.18 on Music-Arena (from 0.70 to 0.88). Since all methods share the same judge, these gains come from rubric construction alone, showing that a better rubric increases agreement with human preferences without changing the judge model. AxisRubric is also competitive with scalar reward models on the benchmarks they were trained on. With the best author-judge configuration, AxisRubric leads or ties all three reward

models on HPS v3, ImageRewardDB, and MJ-Bench, with an average gain of 0.08, through axis discovery and calibration on 40 labeled instances rather than benchmark-specific training (see Appendix B.3). Notably, the improvements also hold across different model choices. Under each of the 21 author-judge configurations in Appendix B.1, AxisRubric achieves higher average accuracy than every baseline on the benchmarks of the corresponding modality (Table 4), and we provide the full results for every configuration in Appendix B.2. To understand the sources of these gains, we conduct an ablation study of each component of AxisRubric in Appendix C. We find that removing any component lowers accuracy in every modality, and that the two largest drops come from removing taxonomy evolution and from removing per-instance rubric generation. We further illustrate how the generated rubric items address axis omission and dilution through two case studies in Appendix F, and provide three examples of complete discovered taxonomies in Appendix E.

Gains Vary Across Benchmarks. Although AxisRubric improves consistently across benchmarks, the size of the gain varies. The smallest gain is 0.02, on HPS v2 (from 0.78 to 0.80) and GenAI-Arena (from 0.68 to 0.70), and the largest is 0.18, on Music-Arena (from 0.70 to 0.88). The gain depends on two factors. The first is the size of the cross-modal observability gap. On HPS v2 and MJ-Bench, preferences mostly follow what the instruction states, so the instruction-derived rubric of TIFA already reaches 0.78 and 0.75, and AxisRubric improves by only 0.02 and 0.03, respectively. In contrast, on HPS v3 and Music-Arena, preferences depend more on delivery quality, such as how well the image is rendered and how the music sounds. The baseline rubrics largely miss these delivery-quality axes, and AxisRubric discovers these axes and gains 0.14 and 0.18, respectively. The second factor is how much of the gap can be expressed in text, since a delivery-quality difference becomes an axis only if the observer can describe it. On Rapidata, both candidate outputs speak the same text, so the gap is entirely about delivery. However, the differences in naturalness and prosody are hard to put into words, so all rubric baselines stay below 0.60. On Music-Arena, the differences include glitches, abrupt endings, and background noise, which the observer can describe precisely, so the gain reaches 0.18. Therefore, AxisRubric gains the most on benchmarks where preferences depend on delivery quality and the relevant differences can be described precisely.

6 Discussion

AxisRubric improves comparative preference accuracy across image, audio, and video generation, but incurs a one-time discovery cost for each benchmark. Discovery is the most expensive stage, since it observes the multimodal outputs over several rounds to build the taxonomy. Once built, the taxonomy is shared by all later comparisons. Hence, this cost is amortized as the evaluation volume grows, and each new comparison requires only rubric generation. Since every rubric item only asks for a quick comparative decision, judging could further be handed to a cheaper System One model such as Jev (Li et al., 2026c) and the multimodal CLM (Kwok et al., 2026), which ranks candidate outputs without generating text.

AxisRubric can be extended in two directions. First, the rubrics can serve as rewards for reinforcement learning. Since the rubric-judge agent answers each rubric item separately, the reward can point to the specific DQ issue an output fails on, and the taxonomy can keep evolving from new outputs as the generator improves. Second, the rubrics can also guide inference-time rejection sampling or best-of-K selection, where per-item verdicts not only identify the better output, but also explain the rationale behind the decision.

7 Conclusion

We presented AxisRubric, a three-phase multi-agent framework that addresses the cross-modal observability gap in per-instance rubric construction for multimodal generation. It systematically discovers preference-relevant evaluation axes and calibrates them against human preferences. Across twelve image, audio, and video benchmarks, AxisRubric improves preference accuracy over the strongest automatic rubric construction baseline in each modality by 0.08 on average and up to 0.18, with consistent gains over 21 rubric-author and output-judge pairs. AxisRubric is thus a practical and generalizable solution for rubric-based preference prediction across diverse modalities. For future work, we plan to use AxisRubric as a reward model for reinforcement learning to improve generative models, and to extend it to rejection sampling and best-of-K selection.

AI USE STATEMENT

In this work, we used generative AI tools for research execution, retrieval and discovery, drafting sections of the paper, and polishing the writing. For research execution, we used AI coding assistants to implement our method and the baselines, and to run the experiments and collect the results. For retrieval and discovery, we used generative AI tools to search for the papers and GitHub repositories of the baseline methods before implementing them. For drafting, we used generative AI tools to organize the collected results into the result tables in LaTeX. For polishing, we used generative AI tools to improve the wording of the paper for readability. We have not used generative AI tools to generate synthetic data, to develop the conceptual framework, to propose or refine hypotheses, to design the methodology or the experiments, or to survey related work. Formulating or proving mathematical claims, translation, and qualitative data analysis are not applicable to this work.

We have reviewed all AI-assisted work. The research questions, the design of our method, the experimental design, and the analysis of the results were all determined by the authors. All AI-generated code was reviewed and verified by the authors, and the baseline implementations were checked against their original papers and public code. The reported results were checked by the authors against the experiment outputs. All AI-polished text was reviewed and revised by the authors, and every claim in the paper was verified by the authors against the experimental results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The source code of our method, including the implementation of all agents, their prompt templates, and the evaluation pipeline, is submitted as supplementary material and will be made public after review. All twelve benchmarks are publicly available, and the sampling of the training and evaluation instances is described in the paper. All models used by the agents are publicly accessible, either through their APIs or as open-weight models, and the models and hyperparameters of every phase are reported in the paper. Note that all agents rely on model APIs, so their outputs may vary across runs. We report the consistency of the judge model across repeated runs, and the reported results are stable under the randomness of API responses.

REFERENCES

Zhaorun Chen, Yichao Du, Zichen Wen, Yiyang Zhou, Chenhang Cui, Zhenzhen Weng, Haoqin Tu, Chaoqi Wang, Zhengwei Tong, Qinglan Huang, Canyu Chen, 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?, 2024. URL https://arxiv.org/abs/2407.04842.

Jaemin Cho, Yushi Hu, Jason Michael 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 *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ITq4ZRUT4a.

Xuelu Feng, Yunsheng Li, Ziyu Wan, Zixuan Gao, Junsong Yuan, Dongdong Chen, and Chunming Qiao. Rubricrl: Simple generalizable rewards for text-to-image generation, 2026. URL https://arxiv.org/abs/2511.20651.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean M. Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=c1bTcrDmt4.

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 *ICCV*, pp. 20349–20360, 2023. URL https://doi.org/10.1109/ICCV51070.2023.01866.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, pp. 21807–21818, 2024. URL https://doi.org/10.1109/CVPR52733.2024.02060.

Ruipeng Jia, Yunyi Yang, Yuxin Wu, Yongbo Gai, Siyuan Tao, Mengyu Zhou, Jianhe Lin, Xiaoxi Jiang, and Guanjun Jiang. Open rubric system: Scaling reinforcement learning with pairwise adaptive rubric. *arXiv preprint arXiv:2602.14069*, 2026a.

Ziheng Jia, Jiaying Qian, Zicheng Zhang, Xiaorong Zhu, Lancheng Gao, and Xiongkuo Min. Raven-eval: Rubric-guided automatic evaluation for ai video generation models based on lmm preference judgement, 2026b. URL https://arxiv.org/abs/2608.09111.

Dongfu Jiang, Max Ku, Tianle Li, Yuansheng Ni, Shizhuo Sun, Rongqi Fan, and Wenhu Chen. Genai arena: An open evaluation platform for generative models, 2024. URL https://arxiv.org/abs/2406.04485.

Zhuoran Jin, Hongbang Yuan, Kejian Zhu, Jiachun Li, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Omni-reward: Towards generalist omni-modal reward modeling with free-form preferences. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=9C4gVbPqSy.

Zijian Kan, Wei Wang, Long Luo, Bing Zhao, Xuan Ren, Weixu Qiao, Wenbo Li, Hu Wei, and Lin Qu. Rubricrm: Generative reward modeling via dynamic rubrics for image generation and editing, 2026. URL https://arxiv.org/abs/2608.26956.

Kuei-Chun Kao, Daixuan Huo, Yuanhao Ban, and Cho-Jui Hsieh. Autorubric-t2i: Robust rule-based reward model for text-to-image alignment, 2026. URL https://arxiv.org/abs/2605.17602.

Yonghyun Kim, Wayne Chi, Anastasios N. Angelopoulos, Wei-Lin Chiang, Koichi Saito, Shinji Watanabe, Yuki Mitsufuji, and Chris Donahue. Music arena: Live evaluation for text-to-music, 2025. URL https://arxiv.org/abs/2507.20900.

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, 2023. URL https://arxiv.org/abs/2305.01569.

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/.

Jacky Kwok, Hangoo Kang, Tarun Suresh, Jon Saad-Falcon, Marco Pavone, Christopher Ré, and Azalia Mirhoseini. Contrastive language models: A system one model for fast and generalizable decision-making, 2026. URL https://contrastive-lm.notion.site. Notion Blog.

Bo Li, Yida Yin, Wenhao Chai, Xingyu Fu, and Zhuang Liu. UEval: A real-world benchmark for unified multimodal generation, 2026a. URL https://openreview.net/forum?id=D4wU9POSMe.

Haitao Li, Tian Tan, Yuguang Yang, Shan Yang, and Xie Chen. Anyaudio-judge: A dynamic rubric-based benchmark and evaluator for audio instruction following, 2026b. URL https://arxiv.org/abs/2606.03116.

Yubo Li, Yidi Miao, Ramayya Krishnan, and Rema Padman. Jev-as-a-judge: Accept when confident, escalate when unsure, 2026c. URL https://arxiv.org/abs/2609.26550.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di ZHANG, Kun Gai, Yujiu Yang, and Wanli Ouyang. Improving video generation with human feedback. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=nHkg4yc7SP.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models, 2024. URL https://arxiv.org/abs/2310.11440.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15086–15095. IEEE, 2025a.

Zi-Ao Ma, Tian Lan, Rong-Cheng Tu, Shu-Hang Liu, Heyan Huang, Zhijing Wu, Chen Xu, and Xian-Ling Mao. T2i-eval-r1: Reinforcement learning-driven reasoning for interpretable text-to-image evaluation, 2025b. URL https://arxiv.org/abs/2505.17897.

Jiaying Qian, Ziheng Jia, Qian Zhang, Zicheng Zhang, Jiayi Guo, Junqi Zhang, Guangtao Zhai, and Xiongkuo Min. Dycorm: Dynamic criterion-aware reward modeling for text-to-image generation, 2026. URL https://arxiv.org/abs/2605.25876.

Rapidata. Text-to-audio human preference benchmark dataset. https://huggingface.co, 2025. Accessed: 2026-09-21.

Juanxi Tian, Fengyuan Liu, Jiaming Han, Yilei Jiang, Yongliang Wu, Yesheng Liu, Haodong Li, Furong Xu, and Wanhua Li. Auto-rubric as reward: From implicit preferences to explicit multimodal generative criteria, 2026. URL https://arxiv.org/abs/2605.08354.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound, 2025. URL https://arxiv.org/abs/2502.05139.

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 *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, Jiayan Teng, Zhuoyi Yang, Wendi Zheng, Xiao Liu, Dan Zhang, Ming Ding, Xiaohan Zhang, Xiaotao Gu, Shiyu Huang, Minlie Huang, Jie Tang, and Yuxiao Dong. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation, 2026. URL https://arxiv.org/abs/2412.21059.

Fangxu Yu, Tao Feng, Dehai Min, Zinan Lin, Weijia Xu, Michael Xu, Philip S. Yu, Ge Liu, and Tianyi Zhou. Reinforcement learning with evolving rubrics as rewards for audio reasoning, 2026. URL https://arxiv.org/abs/2608.02831.

Ziyun Zeng, Zixuan Wang, Yongsheng Yu, Hang Hua, and Jiebo Luo. Videoargus: Agentic rubric-grounded unified evaluation for video generation and editing, 2026. URL https://arxiv.org/abs/2608.05485.

APPENDIX

In the appendix, we include implementation details (Appendix A), full results across all configurations of rubric author and judge models (Appendix B), ablation studies (Appendix C), robustness analyses (Appendix D), discovered taxonomy examples (Appendix E), and case studies (Appendix F).

A SETUP AND REPRODUCIBILITY

In this section, we provide the details of the experimental setup in Section 5, including the benchmarks (Appendix A.1), the baselines (Appendix A.2), the metrics (Appendix A.3), the implementation details (Appendix A.4), and the agent prompt templates (Appendix A.5).

A.1 BENCHMARKS

In this section, we provide details of the twelve benchmarks used in our evaluation. We select six image, three audio, and three video benchmarks that differ in the generation task, the characteristics of the outputs, and the source of human feedback.

Image Generation. We include six image benchmarks that differ in image content, generator quality, and who gives the preference. (i) **Pick-a-Pic** (Kirstain et al., 2023) contains creative images such as characters, fantasy scenes, and artworks in a named style, generated by Stable Diffusion-family models. Its preference is given by the user who requested the images, so the labels reflect what the user asked for rather than what an external annotator considers good. (ii) **HPS v2** (Wu et al., 2023) contains images of both creative content and everyday scenes, generated by nine different text-to-image models and mixed with real photos, with preferences from human annotators. Since the candidate images range from crude to photorealistic, its preferences depend on both instruction following and delivery quality. (iii) **HPS v3** (Ma et al., 2025a) extends HPS v2 to more recent generators such as FLUX and Midjourney, with images of characters, animals, landscapes, architecture, and artworks mixed with real photographs. Its preferences depend heavily on rendering quality, which cannot be determined from the instruction. (iv) **ImageRewardDB** (Xu et al., 2023) contains mostly artworks in a specified style generated by Stable Diffusion, and its preferences come from trained expert annotators who rate each image on alignment, fidelity, and harmlessness before ranking. Its labels thus follow explicit criteria, which tests whether the discovered axes recover the criteria that annotators apply. (v) **MJ-Bench** (Chen et al., 2024) goes further by tying each preference to one of four perspectives: alignment pairs that differ in objects, attributes, counts, or spatial relations, safety pairs where one image contains unsafe content, quality pairs where one image has artifacts such as distorted faces or limbs, and bias pairs that differ in demographic representation. It tests whether the rubric covers concerns such as safety and bias beyond visual appeal. (vi) Finally, **OmniReward-T2I** (Jin et al., 2026) contains images of diverse content generated by recent models, and its preferences are given by annotators under a free-form criterion written for each instance, such as the aspect the comparison should focus on.

Audio Generation. We include three audio benchmarks that cover speech, music, and general sounds. (i) **Rapidata** (Rapidata, 2025) contains speech clips synthesized by different text-to-speech systems, where the two candidate outputs read exactly the same sentence, with preferences from crowd listeners. Since the words are fixed, its preferences depend entirely on delivery, such as naturalness and prosody, so it isolates the delivery-quality part of the observability gap, where no axis can be derived from the instruction. (ii) **Music-Arena** (Kim et al., 2025) moves from speech to music, with clips generated by text-to-music models, both instrumental and with vocals, in the genre, mood, and instrumentation requested by a user, and votes from the same user. Its preferences depend on delivery quality that the request cannot specify, such as glitches, abrupt endings, and background noise. (iii) **OmniReward-T2A** (Jin et al., 2026) contains general sounds such as environmental sounds and sound effects, generated by text-to-audio models. Its preference is given by an annotator who first writes an evaluation criterion for the instance in natural language and then compares the two clips under that criterion.

Video Generation. We include three video benchmarks that differ in who gives the preference and under what criterion. (i) **VideoGen-RewardBench** (Liu et al., 2025) contains short videos of people, animals, objects, and scenes in motion, generated by recent text-to-video models. Human annotators rate each video on visual quality, motion quality, and text alignment, and then give an overall preference. (ii) The video-generation subset of **GenAI-Arena** (Jiang et al., 2024) contains short videos generated by open-source text-to-video models, and the vote is given by the user who made the request. Its user votes complement the annotator labels of VideoGen-RewardBench. (iii) **OmniReward-T2V** (Jin et al., 2026) contains short videos of diverse content generated by recent text-to-video models. For each instance, the annotator writes an evaluation criterion in natural language, such as the aspect the comparison should focus on, and gives the preference under that criterion. This is the same annotation protocol as its image and audio subsets, so the three OmniReward benchmarks allow comparison across modalities.

Data Split. From each benchmark, we sample 40 instances for training and 120 instances for evaluation, after excluding instances with missing media, identical candidate outputs, or tie labels. The training instances are used in axis discovery (Phase 1) and axis weight calibration (Phase 2), while the evaluation instances are used only to evaluate the constructed rubrics.

Benchmarks in the Preliminary Study. The preliminary study in Section 3 uses three benchmarks per modality: HPS v3, ImageRewardDB, and MJ-Bench for image, Rapidata, Music-Arena, and OmniReward-T2A for audio, and VideoGen-RewardBench, GenAI-Arena, and OmniReward-T2V for video. The audio and video benchmarks are all the audio and video benchmarks in our evaluation, and the three image benchmarks are a subset of the six image benchmarks. In Tables 1 and 2, we report the averages over these three benchmarks for each modality.

A.2 Baselines

In this section, we provide the implementation details of the baselines in Section 5.

Holistic and Fixed. **Holistic** has no rubric. The judge directly selects the preferred candidate output without any rubric items. **Fixed** uses one fixed expert-authored rubric for every instance of a modality. For image and video, the rubric contains 17 and 18 criteria adapted from VisionReward (Xu et al., 2026), respectively. For audio, the rubric contains the four axes of Audiobox Aesthetics (Tjandra et al., 2025), namely production quality, production complexity, content enjoyment, and content usefulness.

Decomposition and AnyAudio. **Decomposition** constructs the rubric from the generation instruction alone. It decomposes the instruction into atomic evaluation questions, so the rubric contains only what the instruction states. For audio, we use the instruction-decomposition procedure of AnyAudio-Judge (Li et al., 2026b) and denote this baseline as **AnyAudio**.

TIFA, AutoRubric-T2I, AudioRubrics, and RAVEN. These four methods contain a rubric-construction step followed by other steps, such as model training or downstream generation, and we use only the rubric-construction step. The constructed rubric is then applied under our judging protocol, so these baselines differ from AxisRubric only in how the rubric is constructed. **TIFA** (Hu et al., 2023) generates question-answer pairs from the instruction and uses them to measure image faithfulness. We use these question-answer pairs as rubric items for image, and generate them in the same way for video. **AutoRubric-T2I** (Kao et al., 2026) constructs a rubric from image preference examples and refines it iteratively. We use its rubric construction for image. **AudioRubrics** (Yu et al., 2026) constructs rubrics for audio-reasoning responses and uses them as rewards for policy training. We use its rubric construction for audio and omit the policy training. **RAVEN** (Jia et al., 2026b) constructs a rubric by combining criteria derived from the instruction with basic video-quality criteria. We do not include RubricRM (Kan et al., 2026) and DyCoRM (Qian et al., 2026). Both fine-tune a single model to generate the rubric and predict the preference end-to-end, so the rubric cannot be separated from the trained model and judged by the same judge as the other baselines.

Reward Models. We additionally compare with three reward models for image, PickScore (Kirstain et al., 2023), HPSv2 (Wu et al., 2023), and ImageReward (Xu et al., 2023). Each of them is a scoring model trained on the human preference labels of one image benchmark, namely Pick-a-Pic for PickScore, HPS v2 for HPSv2, and ImageRewardDB for ImageReward. Given the instruction and a candidate output, a reward model outputs a scalar score, and we predict the candidate output with the higher score as the preferred one. These baselines thus use neither a rubric nor a judge.

A.3 METRICS

In this section, we define the two metrics used in Section 5.

Preference Accuracy. We use preference accuracy to measure agreement with human preference labels. For each evaluation instance, the prediction is correct if the predicted preferred output is the same as the human preference label. Preference accuracy is the fraction of correct predictions.

Self-Consistency Rate. We use the self-consistency rate to measure whether the judge returns the same verdicts when asked again. We fix the rubrics and the judge model, and repeat the rubric-item judgments in three rounds. Each round follows the judging protocol in Appendix A.4 and yields one verdict from $\{\mathrm{A}, \mathrm{B}, \mathrm{tie}\}$ per rubric item. Self-consistency rate is the fraction of rubric items whose verdict is identical across all three rounds. Since three rounds can return the same verdict by chance, we additionally report Fleiss' κ across the three rounds, which measures the agreement after removing the agreement expected by chance. A κ close to 1 means the judge returns the same verdicts consistently, and a κ close to 0 means the agreement is no better than chance.

A.4 IMPLEMENTATION DETAILS

In this section, we provide the rubric author and judge models, the hyperparameters, and the judging protocol of AxisRubric.

Models. AxisRubric involves LLMs in all three phases. In all experiments, axis discovery (Phase 1) and axis weight calibration (Phase 2) use Gemini 2.5 Pro, so the calibrated taxonomy of each benchmark is the same for every rubric author and judge. Only the rubric author in Phase 3 and the judge change across experiments. We use Gemini 2.5 Pro, Gemini 3.5 Flash, and Qwen3-Omni-30B-Thinking as both rubric authors and judges, and additionally use Claude Sonnet 5 as a rubric author for image. Since the rubric author and the judge are always different models, there are nine author-judge configurations for image and six each for audio and video. We run every configuration on every benchmark of its modality, which gives ninety results in total, namely nine configurations on each of the six image benchmarks and six configurations on each of the six audio and video benchmarks.

Hyperparameters. We keep all hyperparameters fixed across benchmarks and configurations. All agents use the default decoding settings of their API and a maximum output length of $16,384$ tokens, which is sufficient for the longest outputs. For axis discovery, we set $T_{\max} = 3$ and $\tau_{\mathrm{unmatch}} = 0.05$. We select these values by grid search, and observe that more rounds or a lower threshold add few new axes while increasing the cost of discovery. For axis weight calibration, we set the accuracy threshold $\tau = 0.5$, so an axis earns a weight above 1 only when it predicts the human-preferred candidate output on more than half of the training instances under both presentation orders. We set the deactivation fraction $\rho = 0.5$, so a category is deactivated only when more than half of its axes are below accuracy threshold, which prevents a category from being dropped because of a few noisy axes. We set the maximum multiplier $M = 3$, so the axis weights lie in $[1, 9]$ and the most predictive axis counts at most nine times as much as an axis at or below the threshold. We also select M by grid search. A larger M lets a few axes dominate the vote, and a smaller M leaves the weights too close to uniform for calibration to take effect, both of which lower preference accuracy.

Judging Protocol. Phase 3 in Section 4.3 gives each evaluation instance a rubric, which is a list of rubric items with weights. The rubric-judge agent uses this rubric to decide which of the two candidate outputs is better, and we describe how it makes this decision here. The judge evaluates the

two candidate outputs on each rubric item separately. For each rubric item, the judge returns which candidate output is better on this item. To remove position bias, we ask the judge the same question twice, once with the two candidate outputs in the original order and once in the reversed order. If both answers pick the same candidate output, the rubric item votes for that candidate output. If the two answers differ, we treat the rubric item as a tie, and it contributes no vote to either candidate output. Each vote carries the weight of its rubric item. For each candidate output, we sum the weights of the rubric items that vote for it, and the candidate output with the higher weighted vote sum is predicted as the preferred one. When the two candidate outputs have the same weighted vote sum, we rerun the judge on all rubric items for up to two more rounds. If the sums are still equal after these rounds, we fall back to Holistic judging, where the judge chooses between the two candidate outputs directly without the rubric. As a result, the judge always outputs one preferred candidate output for each evaluation instance, and we compare it with the human preference label to compute preference accuracy.

A.5 Agent Prompt Templates

In this section, we provide the prompt templates of all agents in AxisRubric. For each agent, we summarize its input, instruction, and output format.

Instruction-Blind Observer Agent

Input: Two candidate outputs and the current taxonomy.
Instruction: Report observable delivery-quality differences between the two outputs. Every entry must compare the candidates on one axis. List only axes on which a difference is perceptible, and do not speculate about what was requested. Map each difference to an existing taxonomy axis, or return `null` when no axis fits. Do not predict an overall winner.
Output (JSON): One entry per difference, with `axis`, `difference`, `taxonomy_ref` (or `null`), `label`, `type`, `severity`, and `confidence`. Every entry carries the observation-source tag `src=delivery`.

Instruction-Informed Observer Agent

Input: The generation instruction, two candidate outputs, and the current taxonomy.
Instruction: Report observable differences in requested content and instruction following. Every entry must compare the candidates on one axis supported by the generation instruction. Keep these observations separate from delivery-quality differences. Map each difference to an existing taxonomy axis, or return `null` when no axis fits. Do not predict an overall winner.
Output (JSON): One entry per difference, with `axis`, `difference`, `taxonomy_ref` (or `null`), `label`, `type`, `severity`, and `confidence`. Every entry carries the observation-source tag `src=content`.

Pattern-Discovery and Reconciliation Agents

Input: The observation records, the current taxonomy, and supporting evidence.
Instruction: Group recurring observations into candidate patterns. Revise existing taxonomy nodes using `rename`, `merge`, `split`, `move`, or `delete`. In a separate placement pass, add uncovered patterns supported by at least two observations. Preserve identifiers for retained nodes and maintain observation-source provenance through all edits. Keep observation-source tags distinct from taxonomy-category membership.
Output (JSON): Candidate patterns with supporting observations, typed edits to existing taxonomy nodes, and a separate list of placements for uncovered patterns.

Calibration Agent

Input: The taxonomy axes and the training instances.
Instruction: Assess whether each axis applies to the instance. For each applicable axis, compare the candidates according to its meaning and polarity, and return an A/B verdict. Record axis-level judgments without predicting an overall winner.
Output (JSON): One record per axis, containing its taxonomy reference, applicability decision, and A/B verdict when applicable.

Rubric-Author Agent

Input: The generation instruction and the calibrated taxonomy, including observation-source tags.
Instruction: Determine which active axes apply and rewrite them as concrete evaluation questions. Preserve each axis's meaning and polarity. Do not exclude delivery-quality axes solely because the instruction is silent about them. Apply this protection using the observation-source tag, rather than taxonomy-category membership.
Output (JSON): An applicability decision for each active axis and an evaluation question for each applicable axis, linked to its original taxonomy reference.

Rubric-Judge Agent

Input: Two candidate outputs and the per-instance rubric.
Instruction: Compare the candidates separately for each rubric item. Return A when candidate A is better and B when candidate B is better on that item. Follow the meaning and polarity of each question. Preserve the requested item order and do not aggregate the judgments into an overall winner.
Output (JSON): One A/B verdict per rubric item, in the requested order.

B Full Results Across Rubric Author and Judge Configurations

In this section, we report the results under all 21 author-judge configurations. First, we examine how the accuracy of AxisRubric and the baselines changes when the rubric author and the judge change (Appendix B.1). Second, we report the full results on every benchmark under every configuration (Appendix B.2). Finally, we compare AxisRubric with three reward models trained on image preference data (Appendix B.3).

B.1 Changing the Rubric Author and the Judge Models

In Table 4, we report the average accuracy of AxisRubric over the benchmarks of each modality under each author-judge configuration. Since the strongest baseline may change across configurations, we compare AxisRubric with the baseline that has the highest average accuracy over the same benchmarks in each configuration.

The Improvement of AxisRubric Is Consistent Across Rubric Author and Judge Models. We observe that changing the rubric author and the judge changes the accuracy of AxisRubric by 0.04 on image (0.73 to 0.77), 0.12 on audio (0.63 to 0.75), and 0.08 on video (0.67 to 0.75). Nevertheless, AxisRubric exceeds the strongest baseline in every configuration, with improvements ranging from 0.04 to 0.12 and an average of 0.07. The strongest baseline itself changes with the models, which is **Holistic** in eight configurations (including two ties), **Fixed** in four, and automatic rubric methods in eleven (including two ties). Thus, the best baseline changes with the rubric author and the judge, whereas AxisRubric remains the best regardless of the models. We further find that the variation of AxisRubric comes mainly from the judge rather than the rubric author. With the same judge, changing the rubric author changes audio accuracy by at most 0.06 and video accuracy by at most 0.01, whereas with the same rubric author, changing the judge changes them by up to 0.10 and 0.08. This is because the rubric specifies what the judge should look for, but whether the judge can perceive the specified difference depends on its own ability in that modality. Rubrics from different authors thus lead the same judge to similar accuracy, and a judge with weaker audio or video perception lowers accuracy regardless of who writes the rubric.

Table 4: Average accuracy over the benchmarks of each modality under each author-judge configuration. The strongest baseline is the one with the highest average accuracy in that configuration. Bold marks the highest accuracy per row.

Rubric Author Model	Judge Model	Strongest Baseline	Baseline Accuracy	AxisRubric
Image				
Gemini 2.5 Pro	Gemini 3.5 Flash	TIFA	0.68	**0.74**
Claude Sonnet 5	Gemini 3.5 Flash	AutoRubric-T2I	0.69	**0.74**
Claude Sonnet 5	Gemini 2.5 Pro	Holistic	0.67	**0.77**
Claude Sonnet 5	Qwen3-Omni-30B-Thinking	TIFA	0.68	**0.73**
Gemini 2.5 Pro	Qwen3-Omni-30B-Thinking	AutoRubric-T2I	0.67	**0.73**
Gemini 3.5 Flash	Gemini 2.5 Pro	Holistic	0.67	**0.73**
Gemini 3.5 Flash	Qwen3-Omni-30B-Thinking	AutoRubric-T2I	0.63	**0.75**
Qwen3-Omni-30B-Thinking	Gemini 3.5 Flash	AutoRubric-T2I	0.67	**0.75**
Qwen3-Omni-30B-Thinking	Gemini 2.5 Pro	Holistic	0.67	**0.75**
Audio				
Gemini 2.5 Pro	Gemini 3.5 Flash	Holistic	0.66	**0.73**
Gemini 2.5 Pro	Qwen3-Omni-30B-Thinking	AudioRubrics	0.65	**0.69**
Gemini 3.5 Flash	Gemini 2.5 Pro	Holistic / AnyAudio	0.64	**0.73**
Gemini 3.5 Flash	Qwen3-Omni-30B-Thinking	Holistic	0.59	**0.63**
Qwen3-Omni-30B-Thinking	Gemini 3.5 Flash	Holistic	0.66	**0.71**
Qwen3-Omni-30B-Thinking	Gemini 2.5 Pro	AudioRubrics	0.66	**0.75**
Video				
Gemini 2.5 Pro	Gemini 3.5 Flash	Fixed	0.69	**0.73**
Gemini 2.5 Pro	Qwen3-Omni-30B-Thinking	Holistic / Decomposition	0.58	**0.68**
Gemini 3.5 Flash	Gemini 2.5 Pro	Fixed	0.67	**0.75**
Gemini 3.5 Flash	Qwen3-Omni-30B-Thinking	Decomposition	0.60	**0.67**
Qwen3-Omni-30B-Thinking	Gemini 3.5 Flash	Fixed	0.69	**0.73**
Qwen3-Omni-30B-Thinking	Gemini 2.5 Pro	Fixed	0.67	**0.74**

The Best Judge Model Depends on the Modality. If we compare configurations that share the rubric author model and differ only in the judge model, the effect of the judge model differs across modalities. In Table 4, on image, the judge model changes accuracy by at most 0.04 for any rubric author model, and Qwen3-Omni-30B-Thinking as judge model is as accurate as the two Gemini judge models. With Gemini 3.5 Flash as rubric author model, for example, Qwen3-Omni-30B-Thinking as judge model reaches 0.75, above Gemini 2.5 Pro at 0.73, so a weaker judge model does not hurt image accuracy. On audio and video, however, Qwen3-Omni-30B-Thinking as judge model gives the lowest accuracy for every rubric author model. With Gemini 3.5 Flash as rubric author model, replacing Gemini 2.5 Pro with Qwen3-Omni-30B-Thinking as judge model lowers audio accuracy from 0.73 to 0.63 and video accuracy from 0.75 to 0.67. This is because the rubric only tells the judge model what to look for, and the judge model still has to perceive the difference in the candidate outputs. Qwen3-Omni-30B-Thinking perceives image differences as accurately as the Gemini judge models, but misses more audio and video differences. Therefore, a judge model that works well on image is not necessarily a good judge model for audio or video, and the judge model should be selected separately for each modality.

B.2 Full Results on Every Benchmark

In Tables 5, 6, and 7, we report the accuracy of AxisRubric and all baselines on every benchmark under each of the 21 author-judge configurations, for image, audio, and video, respectively. We find that the strongest baseline differs not only across models but also across benchmarks.

On image, over the 54 results of six benchmarks under nine configurations, the strongest baseline is AutoRubric-T2I in 18 results, Fixed in 16, TIFA in 11, Decomposition in 10, and Holistic in 6, where a result with several tied strongest baselines counts toward each of them. Which baseline is the strongest depends on what the preferences of the benchmark rely on. Fixed is the strongest

in eight of the nine configurations on HPS v2 and in four on ImageRewardDB, where preferences depend on rendering quality that its expert-authored criteria cover. Decomposition is the strongest in six configurations on MJ-Bench, whose comparisons are built around what the instruction specifies, and AutoRubric-T2I is the strongest in six configurations on OmniReward-T2I. On Pick-a-Pic, no baseline dominates, and Holistic is the strongest in three configurations, since preferences depend on the overall visual appeal of the image, which none of the baseline rubrics cover.

The same pattern holds on audio and video. AudioRubrics is the strongest in three of six configurations on Rapidata, since both candidate outputs speak the same text and only a rubric that describes audio properties, rather than the requested content, can separate them. Fixed is the strongest in three configurations on VideoGen-RewardBench and five on OmniReward-T2V, where preferences depend on visual and motion quality that the expert-authored video criteria cover. Holistic is the strongest in three configurations on GenAI-Arena, where the votes reflect the overall impression of the users rather than a specific criterion. Therefore, each baseline is strong only where the axes it covers happen to match what the preferences of the benchmark depend on. In contrast, AxisRubric discovers the axes from the candidate outputs of each benchmark and weighs them by its human preferences, so its rubric adapts to whatever the preferences depend on. As a result, AxisRubric is the best or tied best in 85 of the 90 results, and the five exceptions are within 0.05 of the strongest baseline.

Table 5: Image preference accuracy across nine author-judge configurations. Bold marks the highest accuracy per row and underline marks the second highest, including ties.

Benchmark	Holistic	Fixed	Decomposition	TIFA	AutoRubric-T2I	AxisRubric
Author: Gemini 2.5 Pro; Judge: Gemini 3.5 Flash						
Pick-a-Pic	0.53	0.60	<u>0.70</u>	0.69	0.51	**0.73**
HPS v2	0.73	<u>0.78</u>	0.75	<u>0.78</u>	0.74	**0.80**
HPS v3	0.53	<u>0.70</u>	0.55	0.59	0.64	**0.73**
ImageRewardDB	0.68	<u>0.70</u>	0.65	0.69	0.68	**0.78**
MJ-Bench	0.60	0.58	0.65	<u>0.75</u>	0.61	**0.78**
OmniReward-T2I	0.60	0.60	0.50	0.58	<u>0.62</u>	**0.63**
Average	0.61	0.66	0.63	<u>0.68</u>	0.63	**0.74**
Author: Claude Sonnet 5; Judge: Gemini 3.5 Flash						
Pick-a-Pic	0.53	0.60	0.68	0.68	<u>0.72</u>	**0.73**
HPS v2	0.73	<u>0.78</u>	0.75	0.73	<u>0.78</u>	**0.83**
HPS v3	0.53	**0.70**	0.43	0.63	0.61	<u>0.68</u>
ImageRewardDB	0.68	<u>0.70</u>	0.50	0.58	0.63	**0.75**
MJ-Bench	0.60	0.58	<u>0.73</u>	0.67	0.69	**0.75**
OmniReward-T2I	0.60	0.60	<u>0.67</u>	0.65	**0.68**	**0.68**
Average	0.61	0.66	0.63	0.66	<u>0.69</u>	**0.74**
Author: Claude Sonnet 5; Judge: Gemini 2.5 Pro						
Pick-a-Pic	**0.73**	0.55	0.60	0.58	<u>0.70</u>	**0.73**
HPS v2	0.78	<u>0.85</u>	0.70	0.70	0.70	**0.90**
HPS v3	0.53	<u>0.63</u>	0.53	0.60	0.60	**0.80**
ImageRewardDB	<u>0.64</u>	0.59	0.51	0.56	0.59	**0.72**
MJ-Bench	0.68	0.53	<u>0.70</u>	0.68	0.65	**0.78**
OmniReward-T2I	0.65	0.53	0.60	<u>0.66</u>	0.63	**0.68**
Average	<u>0.67</u>	0.61	0.61	0.63	0.65	**0.77**
Author: Claude Sonnet 5; Judge: Qwen3-Omni-30B-Thinking						
Pick-a-Pic	0.65	0.60	0.58	<u>0.75</u>	0.72	**0.78**
HPS v2	0.75	<u>0.78</u>	0.63	0.73	0.75	**0.88**
HPS v3	<u>0.58</u>	<u>0.58</u>	0.43	**0.63**	**0.63**	**0.63**
ImageRewardDB	0.55	<u>0.65</u>	0.58	**0.73**	0.59	**0.73**
MJ-Bench	0.43	0.45	0.55	0.70	<u>0.73</u>	**0.75**
OmniReward-T2I	0.58	0.58	<u>0.60</u>	0.53	<u>0.60</u>	**0.63**
Average	0.59	0.61	0.56	<u>0.68</u>	0.67	**0.73**
Author: Gemini 2.5 Pro; Judge: Qwen3-Omni-30B-Thinking						
Pick-a-Pic	0.65	0.60	0.55	0.70	<u>0.73</u>	**0.80**
HPS v2	0.75	0.78	0.60	0.78	**0.83**	<u>0.80</u>

Continued on next page.

Image preference results, continued.

Benchmark	Holistic	Fixed	Decomposition	TIFA	AutoRubric-T2I	AxisRubric
HPS v3	0.58	0.58	0.55	0.63	0.60	**0.70**
ImageRewardDB	0.55	0.65	0.58	0.48	0.53	**0.68**
MJ-Bench	0.43	0.45	**0.70**	0.53	0.65	**0.70**
OmniReward-T2I	0.58	0.58	0.48	0.65	0.65	**0.68**
Average	0.59	0.61	0.58	0.63	0.67	**0.73**
Author: Gemini 3.5 Flash; Judge: Gemini 2.5 Pro						
Pick-a-Pic	0.73	0.55	0.55	0.64	0.63	**0.75**
HPS v2	0.78	**0.85**	0.68	0.68	0.60	0.83
HPS v3	0.53	0.63	0.48	0.60	**0.70**	**0.70**
ImageRewardDB	0.64	0.59	0.56	0.51	0.62	**0.69**
MJ-Bench	0.68	0.53	0.68	0.70	0.60	**0.75**
OmniReward-T2I	**0.65**	0.53	0.60	0.58	0.63	0.63
Average	0.67	0.61	0.59	0.62	0.63	**0.73**
Author: Gemini 3.5 Flash; Judge: Qwen3-Omni-30B-Thinking						
Pick-a-Pic	0.65	0.60	0.65	0.55	0.68	**0.83**
HPS v2	0.75	0.78	0.65	0.55	0.73	**0.85**
HPS v3	0.58	0.58	0.60	0.53	0.60	**0.68**
ImageRewardDB	0.55	0.65	0.53	0.55	0.53	**0.75**
MJ-Bench	0.43	0.45	0.63	0.55	0.60	**0.68**
OmniReward-T2I	0.58	0.58	0.63	0.48	0.65	**0.68**
Average	0.59	0.61	0.62	0.54	0.63	**0.75**
Author: Qwen3-Omni-30B-Thinking; Judge: Gemini 3.5 Flash						
Pick-a-Pic	0.53	0.60	0.75	0.65	0.65	**0.83**
HPS v2	0.73	0.78	0.77	0.73	0.74	**0.82**
HPS v3	0.53	**0.70**	0.58	0.65	0.66	**0.70**
ImageRewardDB	0.68	0.70	0.58	0.62	0.72	**0.75**
MJ-Bench	0.60	0.58	0.73	0.65	0.61	**0.75**
OmniReward-T2I	0.60	0.60	0.50	0.61	0.61	**0.65**
Average	0.61	0.66	0.65	0.65	0.67	**0.75**
Author: Qwen3-Omni-30B-Thinking; Judge: Gemini 2.5 Pro						
Pick-a-Pic	0.73	0.55	0.68	0.63	0.68	**0.80**
HPS v2	0.78	**0.85**	0.70	0.65	0.63	**0.85**
HPS v3	0.53	0.63	0.43	0.53	**0.70**	**0.70**
ImageRewardDB	0.64	0.59	0.56	0.56	**0.67**	**0.67**
MJ-Bench	0.68	0.53	0.70	0.65	0.50	**0.78**
OmniReward-T2I	0.65	0.53	0.63	0.66	0.63	**0.68**
Average	0.67	0.61	0.62	0.61	0.64	**0.75**

Table 6: Audio preference accuracy across six author-judge configurations. Bold marks the highest accuracy per row and underline marks the second highest, including ties.

Benchmark	Holistic	Fixed	AnyAudio	AudioRubrics	AxisRubric
Author: Gemini 2.5 Pro; Judge: Gemini 3.5 Flash					
Rapidata	0.57	0.53	0.50	0.48	**0.58**
Music-Arena	0.83	0.78	0.60	0.70	**0.88**
OmniReward-T2A	0.58	0.55	0.63	0.67	**0.73**
Average	0.66	0.62	0.58	0.62	**0.73**
Author: Gemini 2.5 Pro; Judge: Qwen3-Omni-30B-Thinking					
Rapidata	0.48	0.50	0.48	0.52	**0.55**
Music-Arena	0.70	0.70	0.63	0.75	**0.78**
OmniReward-T2A	0.60	0.55	0.58	0.68	**0.75**
Average	0.59	0.58	0.56	0.65	**0.69**
Author: Gemini 3.5 Flash; Judge: Gemini 2.5 Pro					
Rapidata	0.55	0.60	0.65	0.67	**0.70**
Music-Arena	0.80	0.75	0.68	0.65	**0.85**

Continued on next page.

Audio preference results, continued.

Benchmark	Holistic	Fixed	AnyAudio	AudioRubrics	AxisRubric
OmniReward-T2A	0.58	0.50	0.60	0.55	**0.63**
Average	0.64	0.62	0.64	0.62	**0.73**
Author: Gemini 3.5 Flash; Judge: Qwen3-Omni-30B-Thinking					
Rapidata	0.48	0.50	0.56	0.53	**0.58**
Music-Arena	0.70	0.70	0.58	0.63	**0.75**
OmniReward-T2A	**0.60**	0.55	**0.60**	0.55	0.55
Average	0.59	0.58	0.58	0.57	**0.63**
Author: Qwen3-Omni-30B-Thinking; Judge: Gemini 3.5 Flash					
Rapidata	0.57	0.53	0.50	0.55	**0.58**
Music-Arena	0.83	0.78	0.65	0.73	**0.88**
OmniReward-T2A	0.58	0.55	0.60	0.65	**0.68**
Average	0.66	0.62	0.58	0.64	**0.71**
Author: Qwen3-Omni-30B-Thinking; Judge: Gemini 2.5 Pro					
Rapidata	0.55	0.60	0.55	0.67	**0.70**
Music-Arena	0.80	0.75	0.68	0.75	**0.93**
OmniReward-T2A	0.58	0.50	0.60	0.57	**0.63**
Average	0.64	0.62	0.61	0.66	**0.75**

Table 7: Video preference accuracy across six author-judge configurations. Bold marks the highest accuracy per row and underline marks the second highest, including ties.

Benchmark	Holistic	Fixed	Decomposition	TIFA	RAVEN	AxisRubric
Author: Gemini 2.5 Pro; Judge: Gemini 3.5 Flash						
VideoGen-RewardBench	0.63	0.73	0.55	0.50	0.69	**0.75**
GenAI-Arena	0.68	0.63	0.43	0.60	0.68	**0.70**
OmniReward-T2V	0.65	0.70	0.55	0.60	0.65	**0.75**
Average	0.65	0.69	0.51	0.57	0.67	**0.73**
Author: Gemini 2.5 Pro; Judge: Qwen3-Omni-30B-Thinking						
VideoGen-RewardBench	0.51	0.49	0.59	0.54	0.59	**0.69**
GenAI-Arena	0.53	0.53	0.50	0.58	0.50	**0.60**
OmniReward-T2V	0.70	0.70	0.65	0.53	0.60	**0.75**
Average	0.58	0.57	0.58	0.55	0.56	**0.68**
Author: Gemini 3.5 Flash; Judge: Gemini 2.5 Pro						
VideoGen-RewardBench	0.53	0.63	0.65	0.68	0.60	**0.75**
GenAI-Arena	0.68	0.58	0.48	0.50	0.53	**0.70**
OmniReward-T2V	0.63	**0.80**	0.73	0.51	0.65	**0.80**
Average	0.61	0.67	0.62	0.56	0.59	**0.75**
Author: Gemini 3.5 Flash; Judge: Qwen3-Omni-30B-Thinking						
VideoGen-RewardBench	0.51	0.49	0.56	0.59	0.56	**0.64**
GenAI-Arena	0.53	0.53	0.55	0.48	0.53	**0.63**
OmniReward-T2V	0.70	0.70	0.70	0.63	0.55	**0.75**
Average	0.58	0.57	0.60	0.57	0.55	**0.67**
Author: Qwen3-Omni-30B-Thinking; Judge: Gemini 3.5 Flash						
VideoGen-RewardBench	0.63	0.73	0.58	0.64	0.65	**0.75**
GenAI-Arena	0.68	0.63	0.53	0.60	0.69	**0.70**
OmniReward-T2V	0.65	0.70	0.70	**0.75**	0.61	**0.75**
Average	0.65	0.69	0.60	0.66	0.65	**0.73**
Author: Qwen3-Omni-30B-Thinking; Judge: Gemini 2.5 Pro						
VideoGen-RewardBench	0.53	0.63	0.48	0.53	0.60	**0.73**
GenAI-Arena	0.68	0.58	0.53	0.55	0.63	**0.70**
OmniReward-T2V	0.63	**0.80**	0.78	0.63	0.70	**0.80**
Average	0.61	0.67	0.60	0.57	0.64	**0.74**

B.3 Comparison with Reward Models

In Table 8, we compare AxisRubric with three reward models on the six image benchmarks. PickScore, HPSv2, and ImageReward are reward models trained on the human preference labels of Pick-a-Pic, HPS v2, and ImageRewardDB, respectively. We report AxisRubric in two settings. **Primary** uses Gemini 2.5 Pro as rubric author model and Gemini 3.5 Flash as judge model, and **Best** selects the best of the nine author-judge configurations separately for each benchmark.

We first examine how the reward models perform beyond their training benchmarks. PickScore reaches 0.83 on its own training benchmark Pick-a-Pic, but drops to 0.58 to 0.73 on the other five benchmarks, on which the primary configuration of AxisRubric reaches 0.63 to 0.80. Likewise, HPSv2 reaches 0.83 on HPS v2 but only 0.63 to 0.78 elsewhere. This is because each reward model is trained to fit the preference labels of one benchmark, and the other benchmarks differ in image content and in who gives the labels, so what the reward model has learned does not transfer to them. Compared with the reward models, AxisRubric (primary) achieves an average accuracy of 0.74, exceeding HPSv2, the best reward model, by 0.04. On HPS v3, ImageRewardDB, and MJ-Bench, AxisRubric (primary) leads or ties all three reward models. On Pick-a-Pic and HPS v2, AxisRubric (primary) is below PickScore and HPSv2, respectively, which are the reward models trained on these two benchmarks, but still exceeds the other two reward models. On OmniReward-T2I, AxisRubric (primary) is below ImageReward and ties the other two reward models. With the best configuration per benchmark, AxisRubric (best) reaches an average accuracy of 0.80, and matches or surpasses the strongest reward model on all six benchmarks, including the benchmarks that the reward models are trained on. It matches PickScore on Pick-a-Pic, and surpasses HPSv2 on HPS v2 by 0.07 and ImageReward on ImageRewardDB by 0.18. Beyond accuracy, AxisRubric is also interpretable. A reward model outputs only a scalar score and gives no reason for its preference, whereas AxisRubric predicts the preference by weighted voting over rubric items, so every prediction can be traced back to the rubric items and the axes that decided it.

Table 8: Image preference accuracy against reward models. AxisRubric (Primary) uses Gemini 2.5 Pro as rubric author model and Gemini 3.5 Flash as judge model. AxisRubric (Best) reports, for each benchmark, the highest accuracy among the nine author-judge configurations. † marks the benchmark on which the reward model is trained. Bold marks the highest accuracy in each row, including ties.

Benchmark	PickScore	HPSv2	ImageReward	AxisRubric (Primary)	AxisRubric (Best)
Pick-a-Pic	**0.83**†	0.63	0.68	0.73	**0.83**
HPS v2	0.73	0.83†	0.79	0.80	**0.90**
HPS v3	0.58	0.68	0.53	0.73	**0.80**
ImageRewardDB	0.65	0.63	0.60†	**0.78**	**0.78**
MJ-Bench	0.65	**0.78**	**0.78**	**0.78**	**0.78**
OmniReward-T2I	0.63	0.63	**0.68**	0.63	**0.68**
Average	0.68	0.70	0.68	0.74	**0.80**

C Ablation Study

In this section, we examine how much each component of AxisRubric contributes to preference accuracy. We consider four ablations, each of which removes one component of AxisRubric while keeping the rubric author model, the judge model, and the judging protocol unchanged. (i) **Without taxonomy evolution** runs only the first round of axis discovery and uses the taxonomy from that round without further evolution. (ii) **Without consolidation** skips the reconciliation agent in axis discovery, so the axes and the recurring unmatched groups found by the pattern-discovery agent are all kept as they are, without being renamed, merged, split, moved, or deleted. (iii) **Without calibration** skips axis weight calibration, so all categories of the taxonomy stay active and every rubric item receives the same weight. (iv) **Benchmark-level rubrics** skips per-instance rubric generation, so every active axis of the calibrated taxonomy is applied to every instance as a rubric item in its original wording, without checking whether the axis applies to the instance or rewriting it for the instance. In Table 9, we report the results of the four ablations and AxisRubric in the primary configuration on all twelve benchmarks.

Table 9: Ablation results in the primary configuration. Each column reports the average accuracy over the benchmarks of the modality. Bold marks the highest accuracy in each column.

Ablation	Image	Audio	Video
Without taxonomy evolution	0.71	0.64	0.68
Without consolidation	0.71	0.70	0.66
Without calibration	0.73	0.71	0.71
Benchmark-level rubrics	0.68	0.66	0.66
AxisRubric (Ours)	**0.74**	**0.73**	**0.73**

(i) One Round of Axis Discovery Misses Axes. We observe that using the taxonomy from the first round lowers accuracy by 0.03 on image benchmarks, 0.09 on audio benchmarks, and 0.05 on video benchmarks. In the first round, the observer agents identify differences among the candidate outputs, and the pattern-discovery and reconciliation agents evolve the taxonomy from these observations. A single pass of observation, however, captures only part of the differences among the candidate outputs. Each later round observes the training instances again with the evolved taxonomy, identifies differences that the earlier rounds missed, and adds axes for the ones that recur, while also refining the existing axes with the new observations. Stopping after the first round thus leaves these axes undiscovered, which is axis omission. The drop is largest on audio, since audio has more delivery-quality axes to discover among the three modalities, so a single round leaves more preference-relevant evaluation axes undiscovered.

(ii) Unconsolidated Axes Cause Dilution. We find that keeping all axes without consolidation lowers accuracy by 0.03 on the image benchmarks, 0.03 on the audio benchmarks, and 0.07 on the video benchmarks. In this ablation study, the pattern-discovery agent still groups the observations under each axis and groups the recurring unmatched observations, but no agent revises the taxonomy based on these grouped observations. Hence, broad axes are never split, unclear axes are never renamed, overlapping axes are never merged, and every group of recurring unmatched observations becomes a new axis even when an existing axis already covers it. The taxonomy thus accumulates axes that overlap with each other or are vaguely defined. When such a taxonomy is turned into a rubric, overlapping axes produce several rubric items that check the same aspect, so this aspect is counted several times in the weighted vote. Also, vaguely defined axes produce rubric items that do not clearly state what to judge, so the judge model cannot give accurate verdicts on them. Both cases dilute the informative rubric items. The reconciliation agent is therefore necessary to keep one clearly defined axis for each recurring observation.

(iii) AxisRubric Performs Well Even Without Human Preference Labels. We find that removing calibration lowers preference accuracy by 0.01 on the image benchmarks and 0.02 on the audio and video benchmarks. Calibration is the only phase of rubric construction that uses the human preference labels, so this ablation study shows how AxisRubric performs when no labels are available for constructing the rubric. Even without any label, AxisRubric reaches 0.73 on image and 0.71 on audio and video, which still exceeds the average accuracy of every baseline in Table 3, including the baselines that use expert-authored rubrics. The reason is that the axes are discovered from the differences that actually occur among the candidate outputs, so most of them are relevant to human preferences before any label is used, and giving them equal weight loses little accuracy. On top of these axes, calibration further improves accuracy in every modality by deactivating misleading categories and weighing the axes by how well they predict human preference.

(iv) Per-Instance Rubric Generation Is Necessary. We observe that replacing the per-instance rubrics with benchmark-level rubrics lowers accuracy by 0.06 on the image benchmarks and 0.07 on the audio and video benchmarks, which is the largest drop among the four ablations when averaged over the three modalities. In this ablation study, every active axis of the calibrated taxonomy is used directly as a rubric item for every instance, in the same wording as its definition in the taxonomy. This causes two problems. First, an IF axis that the instruction does not raise still becomes a rubric item and still votes, such as a lyric-related axis for instrumental music. Second, a rubric item in the generic wording of the taxonomy does not say what to check in this instance, so the judge model has to guess how the axis applies to the instruction. Both problems dilute the informative rubric items,

which per-instance rubric generation avoids by excluding the inapplicable IF axes and rewriting the remaining axes as concrete rubric items for the instance.

D ROBUSTNESS

In this section, we examine the robustness of AxisRubric from two aspects. First, we check whether the rubrics constructed by AxisRubric allow the judge model to give consistent verdicts when the same rubrics are judged repeatedly (Appendix D.1). Second, we check how much preference accuracy changes when the human preference labels used for axis weight calibration contain errors (Appendix D.2).

D.1 CONSISTENCY OF THE JUDGE MODEL

To test whether the rubrics constructed by AxisRubric allow the judge model to judge consistently, we fix the rubrics of every evaluation instance, run the judge model three times on their rubric items, and measure the self-consistency rate and Fleiss' κ defined in Appendix A.3. We run this test for all nine author-judge configurations, where the three configurations with Claude Sonnet 5 as rubric author model are evaluated on the image benchmarks and the other six configurations on all twelve benchmarks. As shown in Table 10, the self-consistency rate is between 0.88 and 0.94 and the κ is between 0.79 and 0.88 across configurations, with 0.94 and 0.88 in the primary configuration. These results show that the rubrics constructed by AxisRubric allow the judge model to give the same verdict on most rubric items across repeated runs, which in turn keeps the preference predictions stable. The reason is that each rubric item specifies what to compare and what counts as an error under the instruction, so the judge model has a clear question to answer and little room to decide differently from one run to the next. The consistency also holds for every rubric author model and every judge model, so the consistency comes from the rubric items constructed by AxisRubric and does not depend on which models are used.

Table 10: Self-Consistency rate and Fleiss' κ of the judge model with fixed rubrics. Each configuration is evaluated on all benchmarks of the modalities it covers, i.e., the image benchmarks for Claude Sonnet 5 as rubric author model and all twelve benchmarks otherwise.

Rubric Author Model	**Judge Model**	**Self-consistency Rate**	**Fleiss' κ**
Claude Sonnet 5	Gemini 3.5 Flash	0.94	0.88
Gemini 2.5 Pro	Gemini 3.5 Flash	0.94	0.88
Qwen3-Omni-30B-Thinking	Gemini 3.5 Flash	0.91	0.81
Gemini 3.5 Flash	Qwen3-Omni-30B-Thinking	0.92	0.83
Gemini 2.5 Pro	Qwen3-Omni-30B-Thinking	0.91	0.82
Claude Sonnet 5	Qwen3-Omni-30B-Thinking	0.89	0.80
Gemini 3.5 Flash	Gemini 2.5 Pro	0.89	0.81
Qwen3-Omni-30B-Thinking	Gemini 2.5 Pro	0.89	0.81
Claude Sonnet 5	Gemini 2.5 Pro	0.88	0.79

D.2 SENSITIVITY TO CALIBRATION-LABEL NOISE

To examine the robustness of axis weight calibration to incorrect human preference labels, we flip 5% to 40% of the labels of the training instances, recalibrate the axis weights with the flipped labels, and measure how much the preference accuracy drops on each of the ninety benchmark-configuration results. As shown in Table 11, the average accuracy drop grows slowly with the flip rate, from 0.005 with 5% of the labels flipped to 0.008 with 20% and 0.019 with 40%, and the maximum accuracy drop on any single result stays within 0.075. Flipping 40% of the labels makes them almost as bad as random labels, yet the accuracy drops by only 0.019 on average. Therefore, calibration is robust to noise in the training labels. The main reason is that the labels only affect the weights. The axes, the rubric items, and the verdicts of the judge model on them are all obtained without the labels and stay the same, so the flipped labels only change how much each rubric item weighs in the vote. For most instances, the majority of the rubric items favor the same candidate output, so changing the weights does not change the prediction. The prediction changes only for the instances on which the rubric items disagree with each other. On these instances, the weights play a

decisive role in the weighted vote, so some of them are predicted differently under the flipped labels, which leads to the maximum accuracy drop of 0.075. Overall, AxisRubric is robust to noise in the human preference labels.

Table 11: Accuracy drop when a fraction of the calibration labels is flipped. For each of the ninety benchmark-configuration results, we measure how much the accuracy drops, and report the average and the maximum drops over the ninety results.

Labels Flipped	Average Drop	Maximum Drop
5%	0.005	0.050
10%	0.007	0.050
20%	0.008	0.075
30%	0.011	0.075
40%	0.019	0.075

E Discovered Taxonomy Examples

In this section, we provide three full taxonomies discovered by AxisRubric on OmniReward-T2I (Appendix E.1), OmniReward-T2A (Appendix E.2), and OmniReward-T2V (Appendix E.3), as examples for image, audio, and video generation, respectively. For illustration, these three taxonomies are discovered with Qwen3-Omni-30B-Thinking as the model of the observer agents in Phase 1. The taxonomies used in our experiments are discovered with Gemini 2.5 Pro (Appendix A.4), and their axes differ from the examples, but they follow the same three-level form. In this form, a *category* covers a broad area of evaluation, a *subcategory* groups related axes under a category, and an *evaluation axis* states what to compare between the candidate outputs. We list the taxonomies with hierarchical numbering and indentation, and keep the names and definitions exactly as discovered. In Table 12, we summarize the size of each taxonomy.

Table 12: Sizes of the three example taxonomies discovered with Qwen3-Omni-30B-Thinking as the model of the observer agents.

Benchmark	Categories	Subcategories	Evaluation Axes
OmniReward-T2I	5	9	16
OmniReward-T2A	4	8	15
OmniReward-T2V	5	10	21

E.1 Image: OmniReward-T2I

Box E.1: Discovered Image Taxonomy

Model of the observer agents: Qwen3-Omni-30B-Thinking.
Benchmark: OmniReward-T2I.
Hierarchy: Category → Subcategory → Evaluation axis.

1. **Image Quality.** Errors related to the visual fidelity and technical execution of the image generation process, focusing on clarity, texture, color, and artifact presence.
   - 1.1 *Sharpness and Detail.*
     - 1.1.1 **Blurriness.** The image exhibits a lack of sharpness or clarity, resulting in a smudged or blurry appearance of the subject or scene.
     - 1.1.2 **Texture Abstraction.** The image contains overly smooth textures lacking fine detail, such as fabric weave, wood grain, or skin pores, where texture is abstracted or flattened.
   - 1.2 *Color Fidelity.*
     - 1.2.1 **Color Bleeding.** The image shows unnatural color bleeding, blotchiness, or rainbow-like artifacts on surfaces such as glasses or skin, where color gradients are distorted or inconsistent.
     - 1.2.2 **Color Banding.** The image displays visible stepping or color banding in smooth gradients, such as skies or plain walls, resulting in unnatural transitions between colors.

Box E.1: Discovered Image Taxonomy (continued)

2. Structural Integrity. Errors related to the physical and anatomical plausibility of objects and subjects within the scene, focusing on form, proportions, and interactions.
- 2.1 *Anatomical Accuracy.*
  - 2.1.1 **Anatomical Distortion.** The image depicts human or animal subjects with distorted limbs, facial features, or proportions, where the subject appears non-physical or malformed.
- 2.2 *Object Cohesion.*
  - 2.2.1 **Physical Inconsistency.** Objects in the scene have an incoherent physical form, such as floating objects, incorrect interactions, or implausible spatial relationships.
  - 2.2.2 **Shadow and Light Inconsistency.** Shadows and highlights across objects do not originate from a consistent and identifiable light source, resulting in visually conflicting illumination.

3. Spatial and Compositional Integrity. Errors related to the spatial arrangement, framing, and visual balance of the scene, including focus, depth of field, and foreground/background separation.
- 3.1 *Composition and Framing.*
  - 3.1.1 **Composition Imbalance.** The main subject is not clearly identifiable or positioned in a way that creates a visually balanced and engaging frame.
- 3.2 *Focus and Depth.*
  - 3.2.1 **Depth Inconsistency.** The image does not effectively use focus, depth of field, and foreground/background separation to guide the viewer's attention to the intended subject.
  - 3.2.2 **Shading Inconsistency.** The shading on objects' surfaces does not consistently convey depth and volume, resulting in flat or inconsistent 3D representation.

4. Scene and Contextual Fidelity. Errors related to the accuracy of the overall setting, environment, and background details matching the description, including environmental consistency and contextual coherence.
- 4.1 *Scene Consistency.*
  - 4.1.1 **Scene Mismatch.** The scene does not match the description in terms of setting, environment, or background details, resulting in contextual incoherence.
  - 4.1.2 **Object Presence.** The image fails to include specified subjects, objects, or key elements, resulting in missing components from the description.

5. Attribute and Material Fidelity. Errors related to the accuracy of material properties, textures, and attributes of objects, including color, style, and physical characteristics.
- 5.1 *Material Accuracy.*
  - 5.1.1 **Material Misrepresentation.** Textures and materials do not appear appropriate or faithful to the materials they represent, such as wood having no grain, metal lacking reflectivity, or skin without pores.
- 5.2 *Attribute Consistency.*
  - 5.2.1 **Attribute Inconsistency.** The image depicts an entity with incorrect or inconsistent attributes such as material, color, or style, where the properties do not match the description.
  - 5.2.2 **State and Action Mismatch.** The image depicts an entity in an incorrect state, posture, or performing an action that does not match the description.
  - 5.2.3 **Spatial Relationship Error.** The image fails to show correct relative positioning and interaction between two or more entities, resulting in inconsistent spatial relationships.

E.2 AUDIO: OMNIREWARD-T2A

Box E.2: Discovered Audio Taxonomy

Model of the observer agents: Qwen3-Omni-30B-Thinking.
Benchmark: OmniReward-T2A.
Hierarchy: Category → Subcategory → Evaluation axis.

1. Audio Quality. Errors related to the overall fidelity, clarity, and technical quality of the audio signal.
- 1.1 *Frequency Spectrum.*

Box E.2: Discovered Audio Taxonomy (continued)

- 1.1.1 **Narrow Frequency Range.** Audio lacks full spectrum with reduced bass and high frequencies, sounding muffled or tinny.
- 1.1.2 **Muffled Frequency Response.** Audio exhibits reduced high-frequency detail and depth, resulting in a muffled or tinny sound quality.

1.2 *Artifacts and Distortions.*

- 1.2.1 **Synthetic Warbling.** Audio displays "watery" or "warbling" artifacts from synthesis or compression causing unnatural sound quality.
- 1.2.2 **Digital Glitches.** Audio contains glitches, clipping, distortion, or phasing artifacts from digital processing.

1.3 *Background Noise.*

- 1.3.1 **Background Hum.** Persistent low-frequency electronic hum present in background silence.
- 1.3.2 **Background Hiss.** Audible hiss or static in the background that disrupts silence.

2. Sound Authenticity. Errors where sound sources lack realistic timbre, articulation, or natural behavior.

2.1 *Timbre and Realism.*

- 2.1.1 **Synthetic Timbre.** Sound sources exhibit unnatural timbre, warbling, or synthetic qualities instead of authentic realism.
- 2.1.2 **Muddy Clarity.** Poor articulation and definition of sound sources with indistinct clarity.

2.2 *Sound Behavior.*

- 2.2.1 **Abrupt Decay.** Sounds cut off abruptly without natural fade-out or evolution over time.
- 2.2.2 **Unnatural Decay.** Sudden termination of sound events without natural fade-out or decay.

3. Sound Presence and Relationships. Errors related to the correct representation, timing, and interaction of sound events.

3.1 *Sound Presence.*

- 3.1.1 **Missing Sounds.** Requested sound sources are missing or not present as specified in the prompt.
- 3.1.2 **Sound Mismatch.** Sound sources are present but do not match the requested characteristics or qualities.

3.2 *Temporal Relationships.*

- 3.2.1 **Temporal Order Mismatch.** Sound events occur in incorrect sequence or violate the requested timing relationships.
- 3.2.2 **Incorrect Timing.** Sound events are out of sequence or overlap incorrectly, disrupting the requested timing relationships.

4. Listener Comfort. Errors causing discomfort, fatigue, or distraction during listening due to harsh or unnatural audio characteristics.

4.1 *Frequency Discomfort.*

- 4.1.1 **Harsh Frequencies.** Audio contains piercing, fatiguing high frequencies that cause listening discomfort over time.

E.3 VIDEO: OMNIREWARD-T2V

Box E.3: Discovered Video Taxonomy

Model of the observer agents: Qwen3-Omni-30B-Thinking.
Benchmark: OmniReward-T2V.
Hierarchy: Category → Subcategory → Evaluation axis.

1. Visual Fidelity. Errors related to the quality and realism of visual elements within individual frames.

1.1 *Image Quality.*

- 1.1.1 **Blurriness.** Key subjects in the frame are not sharp and well-defined, appearing blurry or pixelated.

Box E.3: Discovered Video Taxonomy (continued)

- 1.1.2 **Texture Distortion.** Textures on surfaces such as fabric, wood, or skin lack detail and appear smeared or plasticky.
- 1.1.3 **Color Artifacts.** Colors within a frame exhibit digital noise, posterization, or color banding.
- 1.1.4 **Flickering.** Rapid, distracting changes in brightness or color occur across frames.

1.2 *Spatial Consistency.*

- 1.2.1 **Motion Artifacts.** Objects exhibit unnatural motion characteristics, such as floating, abrupt stops, or lack of momentum.
- 1.2.2 **Frame Rate Issues.** The rate of motion appears choppy or inconsistent with standard frame rates, resembling a slideshow.
- 1.2.3 **Occlusion Errors.** Objects that reappear after being occluded do not retain their original identity or properties.
- 1.2.4 **Shape Morphing.** Objects change shape or volume in unintended ways, such as "breathing" or morphing into different forms.

2. Temporal Consistency. Errors related to maintaining coherent and plausible behavior over time across frames.

2.1 *Object Identity.*

- 2.1.1 **Attribute Drift.** Visual attributes like color, texture, shape, or size change illogically between frames.
- 2.1.2 **Spontaneous Alteration.** Objects or subjects spontaneously gain extra parts, disappear, or morph into something else illogically.

2.2 *Biomechanics.*

- 2.2.1 **Joint Misalignment.** Limbs bend in unnatural ways or joints appear in incorrect positions during motion.
- 2.2.2 **Interaction Plausibility.** Interactions between objects or characters appear physically implausible.

3. Environmental Consistency. Errors related to the coherence and accuracy of the background environment and setting.

3.1 *Background Stability.*

- 3.1.1 **Background Drift.** Background elements not in motion shift or change position without cause.
- 3.1.2 **Texture Instability.** Background surfaces exhibit unstable textures, such as "boiling" or "simmering" where they should be stable.

3.2 *Setting Fidelity.*

- 3.2.1 **Environmental Misalignment.** The background environment does not match the prompt's description or context.
- 3.2.2 **Composition Issues.** The scene's composition, object placement, or camera angles deviate from the intended setup.

4. Aesthetic and Style Consistency. Errors related to deviations from the intended artistic style, visual aesthetics, and emotional atmosphere.

4.1 *Style Adherence.*

- 4.1.1 **Style Inconsistency.** The video's artistic style or aesthetic changes unexpectedly during the clip.

4.2 *Quality Degradation.*

- 4.2.1 **Quality Drop.** The video's quality degrades over time, with later frames appearing lower quality than earlier ones.

5. Prompt Compliance. Errors related to failing to accurately represent the subjects, objects, and actions specified in the prompt.

5.1 *Object Presence.*

- 5.1.1 **Missing Elements.** Key subjects, characters, or objects described in the prompt are absent from the video.
- 5.1.2 **Incorrect Attributes.** The attributes of subjects or objects do not match the prompt's description.

5.2 *Action Accuracy.*

Box E.3: Discovered Video Taxonomy (continued)

5.2.1 **Action Misrepresentation.** The specified actions or movements are not correctly portrayed.

The three taxonomies show that the discovered axes cover both instruction following and delivery quality. The IF axes check what the instruction asks for, such as *Object Presence* for image, *Missing Sounds* for audio, and *Missing Elements* for video. The DQ axes differ across modalities, since each modality has its own ways of failing. The image taxonomy contains axes such as *Anatomical Distortion* and *Material Misrepresentation*, the audio taxonomy contains *Synthetic Warbling* and *Temporal Order Mismatch*, and the video taxonomy contains *Attribute Drift* and *Flickering*. Each taxonomy provides the benchmark-level axes, from which Phase 3 constructs the rubric of every instance by selecting the applicable axes and rewriting them as rubric items for the instance.

F Case Studies

In this section, we present two case studies to show how axis omission and dilution lead the baselines to the wrong prediction, and how AxisRubric avoids both. We select two instances from HPS v3 on which AxisRubric matches the human preference while Fixed, Decomposition, and TIFA all select the other image (Appendices F.1 and F.2). For each case study, we show the two candidate images and the rubrics of AxisRubric, Fixed, Decomposition, and TIFA. For the rubric of AxisRubric, we also show the axis and the calibrated weight w of each rubric item, and mark in bold the rubric items that decide the prediction.

F.1 Case Study 1: Affectionate Embrace

We compare two images generated for the instruction *"A couple hugs affectionately in a bright, modern room."* As shown in Fig. 3, both images depict a couple embracing indoors, but they differ in how the embrace is rendered. Image A is a portrait of the couple facing the camera, and Image B shows the couple embracing with a kiss on the cheek. Humans prefer Image B. AxisRubric also predicts Image B, whereas Fixed, Decomposition, and TIFA predict Image A. We provide the rubrics of the four methods in Box F.1, and then explain why the baselines fail and AxisRubric succeeds.

Image A **Image B**



Figure 3: Case Study 1. The two candidate images for the instruction *"A couple hugs affectionately in a bright, modern room."* Humans and AxisRubric prefer Image B, whereas Fixed, Decomposition, and TIFA prefer Image A.

Box F.1: Case Study 1: Rubrics of AxisRubric, Fixed, Decomposition, and TIFA

Instruction: "A couple hugs affectionately in a bright, modern room."
Human preference: Image B.

Method: AxisRubric **Prediction: Image B**

Each rubric item is listed with its axis and calibrated weight w. Bold marks the rubric items that decide the prediction.

1. *Missing or Extra Elements* ($w = 2.00$): Is a couple present in the image, and are they depicted inside a room?
2. *Incorrect Visual Properties* ($w = 2.89$): Does the room appear to be bright and have a modern style?
3. *Incorrect Action or Pose* ($w = 2.36$): Is the couple depicted hugging affectionately?
4. *Incorrect Interaction* ($w = 2.36$): Are the two subjects depicted interacting with each other through an affectionate hug?
5. *Incorrect Expression or Gaze* ($w = 1.52$): Do the subjects' facial expressions convey affection?
6. *Incorrect Scene or Setting* ($w = 1.52$): Is the overall scene or setting a bright, modern room?
7. *Incorrect Lighting Style* ($w = 1.52$): Does the lighting in the image align with the request for a 'bright' room?
8. *Incorrect Number of Features* ($w = 1.00$): **Are the human subjects rendered with the correct number of body parts, such as limbs, digits, and facial features?**
9. *Malformed or Fused Anatomy* ($w = 2.59$): **Are the bodies and limbs of the couple well-formed and plausibly arranged, without being distorted or fused, particularly during the hug?**
10. *Unnatural Anatomical Exaggeration* ($w = 1.92$): **Are the features and musculature of the human subjects depicted with realistic proportions?**
11. *Object Fusion or Distortion* ($w = 1.76$): Are distinct objects in the room, like furniture or decor, rendered without being unnaturally merged, distorted, or fused?
12. *Inconsistent Scale or Perspective* ($w = 1.00$): Is the relative size of the couple and the objects in the room consistent with a coherent and realistic perspective?
13. *Implausible Object Form* ($w = 2.03$): Do the objects in the modern room (e.g., furniture, fixtures, decor) have a plausible form, material, and construction?
14. *Lack of Sharpness* ($w = 1.88$): Are the couple and other focal points of the image rendered with crisp, sharp details?
15. *Digital Artifacts* ($w = 1.42$): Is the image free of unwanted visual flaws like glitches, repeating patterns, or watermark-like artifacts?
16. *Inconsistent Light Source* ($w = 1.78$): Are the highlights and shadows on the couple and objects in the room consistent with a coherent light source?
17. *Implausible Reflections* ($w = 1.96$): If there are reflective surfaces in the modern room, do the reflections plausibly correspond to the surrounding environment?
18. *Unnatural Color or Tone* ($w = 2.60$): Does the image have a natural and pleasing overall color and tone, without any distracting or inconsistent color shifts?
19. *Poor Tonal Range* ($w = 2.23$): Does the image exhibit a good tonal range, with detail in both the highlights and shadows, avoiding a flat or washed-out appearance?
20. *Ineffective Lighting or Color Style* ($w = 3.00$): **Does the lighting and color style effectively support the requested mood of an affectionate moment in a bright, modern room?**
21. *Composition* ($w = 2.10$): Is the couple clearly the focal point of a well-balanced and uncluttered composition?
22. *Incoherent Artistic Style* ($w = 1.00$): Does the image maintain a single, cohesive artistic style throughout?
23. *Mismatched Level of Realism* ($w = 2.57$): **Does the image depict the scene with a consistent and plausible level of realism, avoiding an overly artificial or cartoonish appearance?**

Method: Fixed **Prediction: Image A**

1. Is the image symmetrical?
2. Are the objects well-coordinated?
3. Is the main subject prominent?

Box F.1: Case Study 1: Rubrics of AxisRubric, Fixed, Decomposition, and TIFA (continued)

4. Is the image very rich?
5. Is the background beautiful?
6. Is the image very clear?
7. Are the colors bright?
8. Are the colors beautiful?
9. Is the lighting and shadow very distinct?
10. Are the lighting and shadows very beautiful?
11. Are the image details realistic?
12. Are the image details very exquisite?
13. Is the human body in the image completely correct?
14. Is the human face very beautiful?
15. Are the human hands perfect?
16. Can the image evoke a very positive emotional response?
17. Is the image completely safe?

Method: Decomposition **Prediction: Image A**

1. Is the main subject a couple (two people in a romantic relationship)?
2. Are the two people hugging?
3. Is the hug depicted as affectionate?
4. Is the setting an indoor room?
5. Is the room brightly lit?
6. Does the room have a modern style or decor?

Method: TIFA **Prediction: Image A**

1. Who is shown in the image?
2. What are the people doing?
3. What is the lighting like in the room?
4. How is the style of the room described?
5. Where is the scene set?
6. How many people are in the room?

In Case Study 1, we find that the two images differ mainly in DQ rather than in IF. Both images show a couple hugging in a bright, modern room, but Image A renders the overlapping bodies and arms of the couple less plausibly and looks more artificial than Image B. This difference is why humans prefer Image B, and it is also what the baselines fail to capture. Decomposition and TIFA fail because of axis omission. Their rubric items come from the instruction alone and only ask whether a couple, a hug, and a bright modern room are present. Both images satisfy these rubric items, and no rubric item asks how the hug is rendered, so the judge model has nothing to separate the two images and picks Image A. Fixed fails because of dilution. Its rubric does contain rubric items on the human body, face, and hands, but surrounds them with many general aesthetic rubric items, such as whether the image is symmetrical, rich, or colorful, and whether the face is beautiful. Image A is a polished portrait, so it receives the votes of most of these general rubric items, and these votes outweigh the few rubric items about the human body that favor Image B. In contrast, AxisRubric succeeds because it has the axes that capture this difference. It discovers the axes about body structure and realism from the candidate outputs of the benchmark, and rewrites them for this instruction, so that the rubric item about body structure asks specifically about the bodies and limbs of the couple during the hug. These rubric items, shown in bold in Box F.1, favor Image B and decide the prediction.

F.2 Case Study 2: Social Gathering

We compare two images generated for the instruction *"Women enjoy cocktails and food at a social gathering."* As shown in Fig. 4, both images depict women with drinks and food, but they differ in how the food and drinks are rendered. Image A renders the glasses and dishes in the foreground clearly, and Image B shows more participants in the background. Humans prefer Image A. AxisRubric also predicts Image A, whereas Fixed, Decomposition, and TIFA predict Image B. We provide the rubrics of the four methods in Box F.2, and then explain why the baselines fail and AxisRubric succeeds.

Image A



Image B



Figure 4: Case Study 2. The two candidate images for the instruction *"Women enjoy cocktails and food at a social gathering."* Humans and AxisRubric prefer Image A, whereas Fixed, Decomposition, and TIFA prefer Image B.

Box F.2: Case Study 2: Rubrics of AxisRubric, Fixed, Decomposition, and TIFA

Instruction: "Women enjoy cocktails and food at a social gathering."
Human preference: Image A.

Method: AxisRubric **Prediction: Image A**

Each rubric item is listed with its axis and calibrated weight w. Bold marks the rubric items that decide the prediction.

1. *Missing or Extra Elements* ($w = 2.00$): Are all requested elements present — women, cocktails, and food at a social gathering — without extraneous unrequested objects or people?
2. *Incorrect Action or Pose* ($w = 2.36$): Are the women shown actively engaged in enjoying cocktails and food (e.g., holding drinks, eating, or sipping) as the request describes?
3. *Incorrect Interaction* ($w = 2.36$): Do the women appear to be socially interacting with one another, consistent with a group gathering?
4. *Incorrect Expression or Gaze* ($w = 1.52$): Do the women's facial expressions convey enjoyment (e.g., smiling, relaxed, sociable) matching the described social gathering?
5. *Incorrect Scene or Setting* ($w = 1.52$): Does the setting plausibly depict a social gathering where cocktails and food are being enjoyed?
6. *Incorrect Number of Features* ($w = 1.00$): **Do the women have anatomically correct numbers of limbs, fingers, and facial features?**
7. *Malformed or Fused Anatomy* ($w = 2.59$): **Are the women's bodies and faces free of distortion, unnatural fusion, or implausible anatomical arrangement?**

Box F.2: Case Study 2: Rubrics of AxisRubric, Fixed, Decomposition, and TIFA (continued)

8. *Unnatural Anatomical Exaggeration* ($w = 1.92$): **Are the women's proportions and features rendered without unrealistic or exaggerated distortion?**
9. *Object Fusion or Distortion* ($w = 1.76$): **Are the cocktails, food items, and other objects rendered as distinct, undistorted forms without unnatural merging?**
10. *Inconsistent Scale or Perspective* ($w = 1.00$): Are the relative sizes and perspectives of the women, drinks, food, and setting consistent and plausible?
11. *Implausible Object Form* ($w = 2.03$): **Do the cocktails and food items have plausible, realistic forms and materials appropriate to a social gathering?**
12. *Lack of Sharpness* ($w = 1.88$): **Are the women and food/drink items rendered with crisp, in-focus detail?**
13. *Digital Artifacts* ($w = 1.42$): Is the image free of glitches, repeating patterns, or watermark-like artifacts?
14. *Inconsistent Light Source* ($w = 1.78$): Do highlights and shadows across the women, drinks, and setting suggest a single, coherent light source?
15. *Implausible Reflections* ($w = 1.96$): **Do reflections on glossy surfaces such as cocktail glasses plausibly match their surroundings and shapes?**
16. *Unnatural Color or Tone* ($w = 2.60$): Does the image maintain a natural, consistent color tone without distracting or unnatural color casts?
17. *Poor Tonal Range* ($w = 2.23$): Does the image show a balanced tonal range, avoiding blown-out highlights or flat, low-contrast areas?
18. *Ineffective Lighting or Color Style* ($w = 3.00$): Does the lighting and color palette effectively convey the warm, sociable mood of a gathering with cocktails and food?
19. *Composition* ($w = 2.10$): Is the composition balanced and clear, with a discernible focal point among the women, drinks, and food?
20. *Incoherent Artistic Style* ($w = 1.00$): Does the image maintain a single, cohesive artistic style throughout, rather than mixing disparate styles?
21. *Mismatched Level of Realism* ($w = 2.57$): Does the image maintain a consistent and appropriate level of realism throughout the scene?

Method: Fixed **Prediction: Image B**

1. Is the image symmetrical?
2. Are the objects well-coordinated?
3. Is the main subject prominent?
4. Is the image very rich?
5. Is the background beautiful?
6. Is the image very clear?
7. Are the colors bright?
8. Are the colors beautiful?
9. Is the lighting and shadow very distinct?
10. Are the lighting and shadows very beautiful?
11. Are the image details realistic?
12. Are the image details very exquisite?
13. Is the human body in the image completely correct?
14. Is the human face very beautiful?
15. Are the human hands perfect?
16. Can the image evoke a very positive emotional response?
17. Is the image completely safe?

Method: Decomposition **Prediction: Image B**

1. Are the main subjects in the image women?
2. Are there multiple women in the image (rather than just one)?

Box F.2: Case Study 2: Rubrics of AxisRubric, Fixed, Decomposition, and TIFA (continued)

3. Are cocktails present in the image?
4. Is food present in the image?
5. Are the women drinking or holding cocktails?
6. Are the women eating or engaging with food?
7. Does the setting depict a social gathering?

Method: TIFA **Prediction: Image B**

1. Are the people in the image women?
2. Are the women holding cocktails?
3. Is there food present at the gathering?
4. Does the setting depict a social gathering?
5. Are the women enjoying themselves at the event?
6. Are there multiple women in the image?

In Case Study 2, we find that the two images again differ mainly in DQ. Both images show women with cocktails and food at a gathering, but the food and glasses in Image B are blurry and partly merged, whereas those in Image A are sharp and well-formed. This difference is why humans prefer Image A, and it is what the baselines fail to capture. Decomposition and TIFA fail because of axis omission. Their rubric items only ask whether women, cocktails, food, and a social gathering are present, which Image B satisfies more convincingly with its larger crowd, and no rubric item asks whether the food and drinks are rendered well. Without axes about object form, object fusion, sharpness, and reflections, both Decomposition and TIFA pick Image B. Fixed, on the other hand, fails because of dilution. Its rubric does contain two rubric items that could capture the difference, namely whether the image is clear and whether the details are realistic. However, most of its rubric items are general, such as whether the image is rich, whether the background is beautiful, and whether the colors are bright. Image B, with its larger crowd and richer background, wins these general rubric items, and their votes outweigh the two rubric items that favor Image A. In contrast, AxisRubric succeeds because it has the axes that capture this difference. It rewrites the discovered DQ axes for this generation instruction, so that the rubric items in bold in Box F.2 ask specifically whether the cocktails and food are well-formed and sharp and whether the reflections on the glasses are plausible. Image A wins these rubric items, which decide the prediction.

In both case studies, the baselines fail for one of two reasons. Decomposition and TIFA omit the axes on which the two images differ, and Fixed has these axes but dilutes them with general rubric items. AxisRubric avoids both problems by discovering these axes and rewriting them for the instance.