PANOREWARD: ADAPTIVE MULTI-CRITERIA REWARD MODELING FOR PANORAMIC IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Panoramic image generation has become an important component of VR/AR, cinematography, and world modeling. However, progress in this area is constrained by the lack of a reliable reward model, which in turn caps achievable performance. We posit that reward modeling for panoramic generation should be inherently multi-criteria, explicitly integrating multiple evaluation dimensions and adopting adaptive, preference-conditioned weighting to reflect judgment standards across diverse task requirements. To mitigate this bottleneck, we propose PanoReward, a reward model that evaluates panoramic output along three complementary axes: (i) 3D consistency, (ii) polar visual quality, and (iii) image-text consistency. The model is trained on PanoReward-Data, which contains 943,074 task-profile training instances derived from 195,930 source pairs. In panoramic image generation settings, PanoReward exhibits a strong and consistent alignment with human preferences under different task requirements. In addition, PanoReward supports downstream workflows for panoramic generation, *e.g.*, reward-driven post-training. We will release both PanoReward and the associated training dataset.

1 INTRODUCTION

Panoramic image generation aims to map natural-language prompts to visual scenes that can be explored from multiple viewing directions, enabling applications in virtual reality, immersive media, film and television production, and 3D world modeling. Because these applications have different task requirements, the quality priorities of generated panoramas are not identical. Immersive viewing emphasizes perceptual fidelity and visual quality (Chen et al., 2022), text-driven content creation requires faithful semantic and spatial control (Feng et al., 2026), while world modeling demands strong multi-view and 3D consistency for reconstruction and exploration (World Labs, 2025; Team et al., 2026). Consequently, panoramic image generation lacks a universal quality objective, motivating task-dependent optimization across heterogeneous quality dimensions.

The reward model is the primary interface between task specifications and the optimization procedure for generative models. An effective reward model should map task-relevant quality objectives into effective optimization signals. However, existing reward methods for image generation largely depend on generic visual-quality metrics (Zhang et al., 2018; Mittal et al., 2012) or learned evaluators aligned with human perceptual judgment (Wu et al., 2023b;a; Xu et al., 2023; Ma et al., 2025) and semantic consistency (Hessel et al., 2021), and they typically reduce evaluation to a single scalar score used as the optimization objective. Recent studies (Zhang et al., 2024b; Wu et al., 2025; Xu et al., 2026) have started to investigate multidimensional representations of generation quality, yielding more fine-grained reward signals by characterizing different attributes separately. However, these dimensions are commonly aggregated using a fixed combination rule, which imposes a uniform quality objective and limits the capacity to modulate the relative importance of different attributes across tasks. We posit that, in the context of panoramic image generation, the central requirement is not a static notion of "panoramic image quality", but a task-adaptive reward model that can infer quality demands from task instructions and dynamically determine the contribution of each quality dimension to the optimization process.

To address this objective, we propose ***PanoReward***, a task-instruction conditioned, multidimensional reward model for assessing the quality of panoramic image generation along three perspectives: (i) cross-view 3D consistency, (ii) polar visual quality, and (iii) image-text consistency. The

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.

model further modulates the relative contribution of each dimension to the aggregate reward according to the provided task instruction. For model training, we construct ***PanoReward-Data***, a large-scale, high-fidelity, multidimensional preference dataset tailored to instruction-guided panoramic image generation. The dataset comprises 943,074 task-conditioned training instances derived from 195,930 source pairs. It includes outputs produced by state-of-the-art panoramic image generation models (Labs, 2024; Team et al., 2025; Yang et al., 2025b; Feng et al., 2026). To jointly optimize annotation efficiency and label reliability, we employ a two-stage annotation pipeline. First, we pre-rank candidate samples along each quality dimension using vision-language models (Bai et al., 2025), cross-view consistency measures (Asim et al., 2025), and no-reference image quality metrics Mittal et al. (2012), thereby reducing manual annotation effort and mitigating sample noise. Second, human expert annotators review and correct these pre-ranking results under a unified annotation protocol. This produces the final multidimensional preference labels. In addition, to satisfy the demand for preference-sensitive supervision in task-conditioned reward learning, we apply a filtering strategy that prioritizes samples whose preferred outcomes differ across quality-preference settings, thereby increasing the density and informativeness of the supervisory signal for task-profile-aware reward optimization.

To evaluate task-profile-aware reward modeling performance, we additionally develop ***PanoReward-Bench***, a panorama reward evaluation benchmark that is fully independent of the training data and relies exclusively on human annotations. The benchmark is designed to test whether a model can infer and apply the appropriate quality trade-offs under varying task specifications, rather than merely determining whether a panorama is superior according to a single, fixed quality criterion. Trained on PanoReward-Data, PanoReward not only produces task-profile-aware panorama reward estimates but can also be used as a direct feedback signal for optimizing generative models, thereby enabling instruction-dependent optimization toward distinct quality objectives.

On PanoReward-Bench, PanoReward achieves 0.785 micro-averaged pairwise accuracy over 321 non-hard comparisons, with 0.705, 0.857, and 0.778 on 3D consistency, image-text consistency, and polar visual quality, respectively. On 180 G/T trade-off pairs, learned task-profile-aware fusion achieves 0.867 flip-both accuracy, with 0.917 under G-priority instructions and 0.900 under T-priority instructions. Ablation studies show that task-conditioned preference supervision and learned fusion are both important for instruction-dependent preference following. PanoReward further generalizes to unseen panorama outpainting settings with 0.776 overall pairwise accuracy, and its use as a frozen reward enables task-dependent quality trade-offs in DiT360 post-training. These results demonstrate the effectiveness of PanoReward as a task-profile-aware evaluator and optimization signal for panoramic generation.

2 Related Works

Panoramic Image Generation. Recent methods for panoramic image generation have transitioned from panorama-specific adaptations of diffusion models to geometry-aware and world-generation paradigms. Early approaches (Chen et al., 2022; Tang et al., 2023; Zhang et al., 2024a; Kalischek et al., 2025) primarily exploit pretrained image-generation models while explicitly enhancing global continuity and cross-view consistency. Subsequent work by Yang et al. (2025a;b); Team et al. (2025); Lu et al. (2025) more directly links panoramic synthesis to spatially navigable, explorable world generation. More recently, DiT360 (Feng et al., 2026) employs hybrid perspective-panorama training to improve panoramic fidelity, whereas HY-World 2.0 (Team et al., 2026) further incorporates text-image-to-panorama generation within a broader world-generation pipeline. Despite these developments, the literature has predominantly emphasized architectural and training improvements to the generative model itself, with comparatively limited attention to reward-based mechanisms to optimize panoramic generation models.

Evolution of Image Generation. Image generation reward modeling has evolved from generic perceptual metrics to human-aligned, multidimensional, and increasingly context-aware reward functions. Early metrics such as LPIPS (Zhang et al., 2018) and NIQE (Mittal et al., 2012) primarily measure perceptual or low-level visual quality, while HPS (Wu et al., 2023b;a; Ma et al., 2025), ImageReward (Xu et al., 2023), PickScore (Kirstain et al., 2023), and TIFA (Hu et al., 2023) introduce human-preference or semantic-alignment signals for evaluating and optimizing generation. Recent works (Zhang et al., 2024b; Xu et al., 2026) extend reward modeling beyond a single holis-

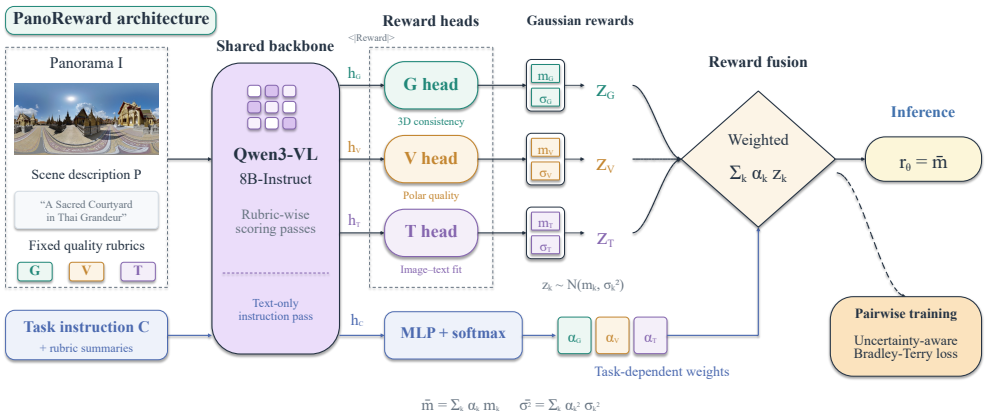


Figure 1: **PanoReward architecture.** The finetuned Qwen3-VL backbone produces rubric-specific G/V/T scores without the task instruction. A text-only branch maps the instruction and rubric summaries to fusion weights α; the scalar reward is $r_\theta = \sum_k \alpha_k m_k$. Training uses the uncertainty-aware Bradley-Terry objective.

tic score by explicitly modeling multiple quality dimensions, while EditReward (Wu et al., 2025) demonstrates how high-quality task-profile-aware preference data can support reward-based optimization. More recently, Wang et al. (2026); Lee et al. (2026) explore adapting reward criteria to individual preferences or contextual requirements. These studies establish a clear trend toward richer and more adaptable reward modeling, but existing approaches primarily target generic image preferences, personalized preferences, or editing-specific objectives. In contrast, PanoReward focuses on task-profile-aware reward modeling for panoramic generation, dynamically adjusting the contribution of panorama-specific quality dimensions, including visual quality, semantic alignment, and 3D consistency, according to the intended task.

3 PANOREWARD

PanoReward is a task-profile-aware reward model for panoramic generation. Given a panorama I, scene description P, and task instruction C, it predicts a scalar reward $r_\theta(I, P, C)$. We separate dimension measurement from task preference: three rubric-specific heads score 3D consistency (G), polar visual quality (V), and image-text consistency (T), while a text-only branch maps C to weights that combine these scores.

3.1 MULTI-DIMENSIONAL SCORING

We use a pretrained Qwen3-VL backbone Bai et al. (2025) with three parallel heads. For dimension k, the head reads the <|Reward|> representation and predicts $(m_k, \ell_k) \in \mathbb{R}^2$, where $\sigma_k = \exp(\ell_k)$ is the predicted scale used by the uncertainty-aware loss. The heads receive I and P, but not C, and therefore produce instruction-agnostic dimension scores.

3.2 TASK-PROFILE-AWARE REWARD FUSION

The task-profile-aware fusion branch encodes the three rubric summaries and C at a special <|Reward|> token. A lightweight MLP and softmax produce $\alpha \in (0,1)^3$ with $\sum_k \alpha_k = 1$. The scalar reward is $r_\theta(I, P, C) = \sum_k \alpha_k m_k$, **e.g.**, instructions prioritizing G or V therefore increase the corresponding fusion weight.

Discussion. The fixed rubric heads make dimension scores comparable across tasks, while the fusion branch changes only their relative weights. Zero initialization gives uniform weights at the start of training. With $z_k \sim \mathcal{N}(m_k, \sigma_k^2)$ and independent dimension outputs, the fused reward has mean $\bar{m} = \sum_k \alpha_k m_k$ and variance $\bar{\sigma}^2 = \sum_k \alpha_k^2 \sigma_k^2$. We use $\bar{m}$ as the reward and $\bar{\sigma}$ in the uncertainty-aware loss. Without an instruction, uniform weights define an instruction-free baseline.

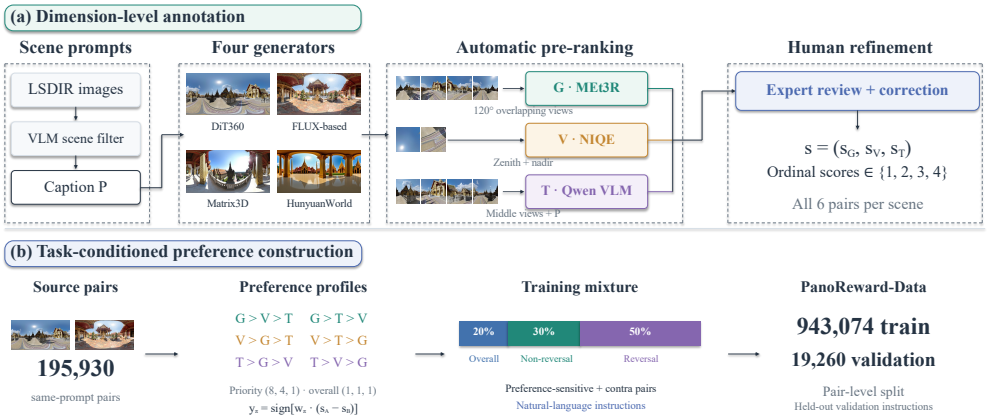


Figure 2: **PanoReward-Data construction.** (a) Four generators produce same-prompt candidates. MEt3R, NIQE, and Qwen3-VL provide initial G/V/T rankings, which trained annotators review into ordinal scores. (b) The 195,930 source pairs are expanded with six priority profiles using $(8, 4, 1)$ and one uniform profile. After filtering and data balancing, it can be split into 943,074 training and 19,260 validation instances.

3.3 Training Objective

We train PanoReward using task-conditioned pairwise preference data constructed from the refined multidimensional quality scores. For each non-tie preference pair, let (z^+, z^-) denote the fused reward variables for the chosen and rejected panoramas. Under the Gaussian parameterization, $z = z^+ - z^-$ has mean $\mu_z = m^+ - m^-$ and standard deviation $\sigma_z = \sqrt{(\sigma^+)^2 + (\sigma^-)^2}$. Following an uncertainty-aware Bradley-Terry model, we minimize

$$\mathcal{L}_i = -\mathbb{E}_{\epsilon\sim\mathcal{N}(0,1)} \left[\log \text{sigmoid}(\mu_z + \sigma_z \epsilon)\right].$$

The task instruction is provided only to the fusion branch.

4 PanoReward-Data Curation

4.1 Dataset Construction

Training a task-profile-aware reward model requires preference supervision that captures both the overall quality of a generated panorama and the relative quality of different dimensions. We therefore construct PanoReward-Data from a text-to-panorama generation pipeline. The source corpus contains 195,930 source pairs, where each pair consists of two panoramas generated from the same scene description by two different state-of-the-art panorama generation models. Each sample consists of a pair of panoramas together with three dimension-wise ordinal scores for each panorama. The dimension scores are first obtained from automatic evaluators and subsequently refined by trained human annotators. We use these refined dimension-level annotations to construct task-conditioned pairwise preferences.

Regarding the three panorama quality dimensions: 3D consistency (G), polar visual quality (V), and image-text consistency (T), we first employed a set of automated ranking collection schemes. Among them, 3D consistency measures the stability of scene structure across viewing directions using a multi-view consistency metric (Asim et al., 2025) on the overlapping middle views of the panorama, polar visual quality measures artifacts and distortions around the zenith and nadir using NIQE (Mittal et al., 2012), and image-text consistency measures the semantic agreement between the generated panorama and the scene description using Qwen3-VL (Bai et al., 2025). Each dimension is represented by an ordinal score in $\{1, 2, 3, 4\}$, where a higher value indicates better quality. The resulting rankings are subsequently refined by trained human annotators following a standardized protocol, yielding the final ordinal scores. Thus, the automatic evaluators provide an efficient initial ordering, while human refinement determines the final dimension-level annotations. Impor-

tantly, these dimension scores are independent of the downstream task and remain unchanged across different task conditions.

4.2 Task-Conditioned Preference Construction

To represent different task requirements, we define seven task profiles. Six profiles specify a strict ordering of dimension priorities,

$$\Pi_{\text{task}} = \{V > T > G,\, V > G > T,\, T > V > G,\, T > G > V,\, G > V > T,\, G > T > V\}, \quad (1)$$

while the seventh profile, overall, assigns equal importance to all three dimensions.

For each strict profile, an integer-programming search selects the fixed priority-weight template $(8, 4, 1)$ by maximizing informative preference-sensitive and reversal pairs subject to the tie-rate and ordering constraints. This search does not use validation or benchmark data. The weights are used only to construct labels. The model receives the corresponding natural-language instruction. For example, the profile $T > V > G$ corresponds to $w = (1, 4, 8)$. The weights are used only during data construction and are never exposed to the reward model. The model receives only the corresponding natural-language task instruction.

Given a pair of panoramas I_A and I_B with dimension scores $s^{(A)}, s^{(B)} \in \{1, \ldots, 4\}^3$, the task-conditioned preference label under profile π is

$$y_\pi = \text{sign}\left(w(\pi)^\top s^{(A)} - w(\pi)^\top s^{(B)}\right) \in \{-1, 0, +1\}. \quad (2)$$

Samples resulting in $y_\pi = 0$ are discarded for that profile. For the overall profile, we use uniform weighting, $w(\text{overall}) = (1, 1, 1)$, and derive the pairwise label directly from the refined dimension scores.

A source pair is considered *preference-sensitive* if there exist two task profiles π_1 and π_2 such that $y_{\pi_1} \neq y_{\pi_2}$. Among these samples, a *preference reversal* occurs when the winner under a task-specific profile differs from the winner under the overall profile: $y_\pi \neq y_{\text{overall}}$. These reversal pairs provide the most informative supervision for learning a task-profile-aware reward, since the same pair of panoramas receives different labels under different task requirements. We define sensitivity and reversal only over profiles whose labels are non-ties and are retained as training instances.

4.3 Data Balancing

Each profile has 10 training templates and 2 held-out validation paraphrases. The numerical weights are omitted from all instructions. The training mixture contains 20% overall-profile, 30% non-reversal counterfactual, and 50% reversal instances, with limited resampling for balance. Splits are made at the source-pair level, yielding 943,074 training instances and 19,260 validation instances.

We additionally identify *contra* cases, in which the winner determined by the highest-priority dimension alone disagrees with the final winner under the full weighted combination. Such samples cannot be solved by relying exclusively on the top-ranked dimension and therefore encourage the model to integrate evidence from multiple quality dimensions. We upsample these cases during training while reducing redundant cases whose labels can already be determined by the highest-priority dimension.

Finally, train/validation splits are performed at the original pair level, such that all task-conditioned variants derived from the same source pair are assigned to the same split. The validation set uses held-out instruction templates to avoid both pair-level leakage and template-level memorization. The resulting PanoReward-Data contains 943,074 training instances and 19,260 validation instances.

5 Experiments

We evaluate PanoReward through four aspects. We first measure agreement with human preferences on PanoReward-Bench. We then evaluate task-profile following under conflicting quality requirements. Next, we test PanoReward as a frozen reward model for reinforcement-learning post-training

Table 1: Inter-annotator agreement on PanoReward-Bench under the final annotation protocol. α uses ordinal distance.

Dimension	Fleiss' κ	Krippendorff's α
3D consistency (G)	0.511	0.687
Image-text consistency (T)	0.454	0.799
Polar visual quality (V)	0.485	0.694

of DiT360. Finally, we conduct ablations to isolate the contributions of reward aggregation and task-conditioned supervision.

5.1 Experimental Setup

5.1.1 Implementation Details

PanoReward uses Qwen3-VL-8B-Instruct with three scoring heads and a zero-initialized two-layer fusion MLP with hidden size 32. We train for one epoch with learning rate 2×10^{-6}, 5% warm-up, per-device batch size 1, gradient accumulation 4, and bfloat16 precision on 8 GPUs. Further implementation details and the complete input templates are provided in Appendix D.

5.1.2 PanoReward-Bench and Metrics

PanoReward-Bench is a multi-way preference benchmark for panoramic image reward models. Each case contains one scene description and four candidate panoramas generated under the same description by different panorama systems. For each case, annotators evaluate three quality dimensions: 3D consistency (G), polar visual quality (V), and image-text consistency (T). We call each case and dimension combination a *cell*. Each cell contains one *4-way ranking*, which means that the four candidates are ordered from best to worst.

We use 10 expert annotators and perform per-cell adjudication following the procedure described in Appendix A. Candidate identities are anonymized and the slot assignment is independently randomized for every case. For each cell, we construct a human consensus ranking by aggregating the annotator rankings. For each candidate, we sum its ranks across annotators, with rank 1 denoting the most preferred candidate, and order candidates by the resulting rank sum. A cell is called *hard* when the rank-sum gap between the best and second-best candidate is smaller than 2. Cells that do not satisfy this ambiguity criterion are called *non-hard cells* and form the main evaluation set. The resulting main set contains 35 G cells, 42 T cells, and 30 V cells.

For the main evaluation, we convert each non-hard 4-way ranking into three pairwise comparisons. In each comparison, the human-preferred panorama is paired with one of the other three candidates. The model selects the panorama with the higher reward score, and the prediction is correct when it agrees with the human preference. This produces 105, 126, and 90 comparisons for G, T, and V, respectively, for 321 comparisons in total. Each comparison receives one point for a correct winner, zero for an incorrect winner, and 0.5 for a score tie. The *micro-averaged* accuracy pools all comparisons with equal weight, so G, T, and V contribute to the All score in proportion to their comparison counts.

We additionally identify 180 *trade-off pairs* for evaluating task-profile following. A trade-off pair consists of two panoramas from the same scene whose human preferences differ between G and T. Specifically, the human G ranking prefers one panorama while the human T ranking prefers the other. These pairs create an explicit conflict between 3D consistency and image-text consistency and therefore provide a direct test of whether a reward model changes its preference according to the task instruction.

5.1.3 Baselines

We compare PanoReward with general VLM judges. GLM-5.3-Flash, Gemini-3.5-Flash, GPT-5.5, Qwen3-VL-8B, and Qwen3.5-9B are evaluated with the same dimension rubrics and task instructions used for PanoReward. We also report leave-one-annotator-out human agreement as a reference.

Table 2: Preference agreement and task-profile following. G, T, V, and All report pairwise accuracy on 105, 126, 90, and 321 comparisons, respectively. Flip-both, G inst., and T inst. are evaluated on 180 G/T trade-off pairs. Random pairwise accuracy is 0.500 and random flip-both accuracy is 0.250.

Method	321 Pairwise Comparisons				180 Trade-off Pairs		
	G	T	V	All	Flip-both	G inst.	T inst.
Random	0.500	0.500	0.500	0.500	0.250	0.500	0.500
GLM-5.3-Flash	0.276	0.754	0.583	0.550	0.011	0.044	0.642
Gemini-3.5-Flash	0.362	0.793	0.572	0.590	0.028	0.161	0.687
GPT-5.5	0.362	0.810	0.539	0.587	0.036	0.245	0.791
Qwen3-VL-8B	0.471	0.571	0.506	0.520	0.000	0.017	0.261
Qwen3.5-9B	0.329	0.659	0.589	0.531	0.006	0.100	0.478
PanoReward + fixed $(8, 4, 1)$	0.686	0.722	0.722	0.710	0.822	0.867	**0.906**
PanoReward (ours)	**0.705**	**0.857**	**0.778**	**0.785**	**0.867**	**0.917**	0.900
Human rankings	0.801	0.862	0.816	0.829	-	-	-



R ↑: task-conditioned reward. Parentheses: reference rank ↓. Bold: highest R within each profile.
G: 3D geometric consistency V: polar visual quality T: image–text consistency
G/T reference: human Borda ranking. V reference: human + NIQE. Rank 1 is best.

Figure 3: Task-conditioned rewards and reference ranks for four generators on one restaurant scene. Each cell reports the reward under the indicated priority profile and the corresponding rank.

5.2 Main Results: Preference Agreement and Task-Profile Following

Table 2 combines the two complementary evaluations. The first four columns measure pairwise agreement on the 321 main comparisons. The last three columns measure task-profile following on the 180 trade-off pairs. *Flip-both* is the fraction of trade-off pairs for which both instruction-specific decisions agree with the corresponding human preferences. G inst. and T inst. denote accuracy under the G-priority and T-priority instructions, respectively.

On the 321 pairwise comparisons, PanoReward achieves 0.785 micro-averaged accuracy, substantially above all general VLM judges and close to the human reference of 0.829. It reaches 0.705 on G, 0.857 on T, and 0.778 on V. The largest improvement over the VLMs appears on G, where PanoReward exceeds the strongest VLM by more than 0.11. These results indicate that the learned reward scores better capture the dimension-specific human preferences.

The trade-off evaluation further tests whether the model changes its preference when the task changes. PanoReward with fixed $(8, 4, 1)$ weights achieves 0.822 flip-both accuracy, while learned fusion reaches 0.867. PanoReward also achieves 0.917 and 0.900 under the G-priority and T-priority instructions, respectively. In contrast, the general VLM judges obtain flip-both accuracies below 0.04. Thus, strong instruction following does not emerge reliably from generic VLM judging alone, while the task-conditioned reward model consistently adapts its preference to the requests.

5.3 Ablation Studies

We run two ablations. First, we compare different aggregation rules while keeping the task-conditioned dimension heads fixed. Second, we remove task-conditioned supervision during training. Table 3 shows these experiments. Detailed head-correlation analysis is provided in Appendix F.

Table 3: Ablation results on the 321 pairwise comparisons and 180 trade-off pairs.

Variant	G	T	V	All	Flip-both
No task-profile supervision + uniform	0.324	0.770	0.700	0.604	0.000
Task-profile supervision + uniform	0.381	**0.865**	0.700	0.660	0.000
Task-profile supervision + fixed $(8, 4, 1)$	0.686	0.722	0.722	0.710	0.822
Task-profile supervision + learned fusion	**0.705**	0.857	**0.778**	**0.785**	**0.867**

Aggregation. With the same task-conditioned heads, overall pairwise accuracy increases from 0.660 with uniform aggregation to 0.710 with fixed $(8, 4, 1)$ weights and 0.785 with learned fusion. Flip-both accuracy follows the same trend, increasing from 0.000 to 0.822 and then to 0.867. This shows that learned fusion improves both general preference agreement and instruction-dependent preference changes.

Task-conditioned supervision. Removing task-conditioned supervision reduces the overall pairwise accuracy to 0.604 when fixed aggregation $(8, 4, 1)$ is used and yields only 0.022 flip-both accuracy. In contrast, task-conditioned supervision with learned fusion reaches 0.785 overall pairwise accuracy and 0.867 flip-both accuracy. These results show that the training signal is necessary for learning dimension-specific reward that can support instruction-dependent preferences.

5.4 Reward-Driven Post-Training

We evaluate whether PanoReward can provide a useful optimization signal for panoramic image generation. We use PanoReward as a frozen reward model for reinforcement-learning post-training of DiT360 (Zheng et al., 2026; Feng et al., 2026). We train one policy per reward condition. The *overall* condition weights the three dimensions equally; priority conditions (e.g., $G > V > T$) are weighted by PanoReward.

Evaluation Metrics. Following DiT360 (Feng et al., 2026), we report panorama fidelity, perceptual quality, and image-text alignment. FAED (Oh et al., 2022) measures distributional distance in panorama autoencoder feature space; IS (Salimans et al., 2016) measures semantic recognizability and diversity; CS uses CLIP (Hessel et al., 2021) for image-text alignment. $\text{QA}_{\text{quality}}$ and $\text{QA}_{\text{aesthetic}}$ use Q-Align (Wu et al., 2024) to measure perceptual and aesthetic quality. FID_{clip}, FID_{pole}, and FID_{equ} compute distributional distance under different panorama evaluation settings (Heusel et al., 2017; Sun et al., 2025). Lower is better for FAED and FID; higher is better for IS, CS, and Q-Align.

Results. The overall reward condition improves all eight reported metrics over the base DiT360. FAED decreases from 2.91 to 2.86, while IS, CS, $\text{QA}_{\text{quality}}$, and $\text{QA}_{\text{aesthetic}}$ increase. FID_{clip}, FID_{pole}, and FID_{equ} also decrease from 41.60, 50.88, and 24.77 to 40.30, 49.10, and 24.10, respectively. This shows that using PanoReward as an optimization signal can improve the overall quality of the generated panoramas rather than only changing their task-specific trade-offs.

The priority conditions produce distinct metric profiles. With $G > V > T$, the model achieves the lowest FAED of 2.79 and the highest IS of 1.67 among all reported models. However, its FID_{pole} and FID_{equ} are higher than those of the overall model, indicating that improvements are not uniform across all panorama quality measures. With $V > T > G$, the model obtains the highest $\text{QA}_{\text{quality}}$

Table 4: DiT360 post-training with frozen PanoReward on held-out prompts. The overall condition gives equal importance to the three dimensions. Priority conditions assign descending importance according to the indicated order.

Model	FAED↓	IS↑	CS↑	$QA_{quality}$↑	$QA_{aesthetic}$↑	FID_{clip}↓	FID_{pole}↓	FID_{equ}↓
DiT360 (base)	2.91	1.60	34.68	4.69	4.19	41.60	50.88	24.77
+ RL (overall)	2.86	1.62	35.42	4.82	4.33	40.30	49.10	24.10
+ RL ($G > V > T$)	2.79	1.67	35.40	4.76	4.20	42.10	52.30	25.80
+ RL ($V > T > G$)	2.92	1.55	34.52	4.87	4.44	39.20	46.80	22.40
+ RL ($T > V > G$)	2.95	1.53	36.25	4.72	4.16	38.30	54.20	26.70

of 4.87 and $QA_{aesthetic}$ of 4.44, together with the lowest FID_{pole} of 46.80 and FID_{equ} of 22.40. In contrast, the $T > V > G$ condition achieves the highest CS of 36.25 and the lowest FID_{clip} of 38.30, while showing larger FAED, FID_{pole}, and FID_{equ} values. These results demonstrate that task-conditioned rewards can steer DiT360 toward different quality trade-offs.

5.5 Out-of-Distribution (OOD) Generalization

We evaluate PanoReward on OOD panorama outpainting settings that are not present in the training source pairs. For each setting, we construct 20 scenes with four candidates per scene and obtain 51 G, 57 T, and 39 V dimension-wise reference preferences. We use the same pairwise evaluation protocol and tie handling as in the main benchmark. Table 5 reports pairwise accuracy for G, T, V, and their micro-average All.

Table 5: Pairwise accuracy on OOD panorama outpainting task. We report G, T, V, and the micro-average All. The same reference protocol and tie handling are used as on PanoReward-Bench.

	Outpainting			
Method	G	T	V	All
GLM-5.3-Flash	0.353	0.851	0.590	0.609
Gemini-3.5-Flash	0.402	0.825	0.513	0.595
GPT-5.5	0.412	0.851	0.513	0.609
Qwen3-VL-8B	0.471	0.553	0.526	0.517
Qwen3.5-9B	0.382	0.632	0.500	0.510
PanoReward	**0.706**	0.825	**0.795**	**0.776**

PanoReward achieves the highest overall accuracy, reaching 0.776 compared with 0.609 for the strongest baseline. The advantage is particularly clear on G and V, where PanoReward reaches 0.706 and 0.795, respectively. These results exceed the strongest baseline by 0.294 on G and 0.205 on V. On T, PanoReward achieves 0.825, which is slightly below the best VLM judges at 0.851. Nevertheless, PanoReward maintains the strongest overall agreement across the three dimensions. This result shows that the learned reward model generalizes to the held-out outpainting setting.

6 Conclusion

We present PanoReward, an instruction-conditioned reward model that separates dimension-specific scoring from task-dependent aggregation for panoramic image generation. PanoReward-Data provides 943,074 training samples derived from multidimensional annotations. On PanoReward-Bench, PanoReward achieves 0.785 pairwise accuracy over 321 human preference comparisons and 0.867 flip-both accuracy on 180 G/T trade-off pairs. Ablation studies show that both task-conditioned supervision and learned reward fusion are important for reliable instruction-dependent preference following. PanoReward also generalizes to unseen panorama outpainting settings, achieving 0.776 overall pairwise accuracy. When used as a frozen reward for DiT360 post-training, PanoReward improves the overall generation quality and produces distinct metric trade-offs under different task priority profiles, demonstrating its potential as a controllable optimization signal for panoramic image generation. PanoReward-Bench remains limited in scene and generator coverage, and broader evaluation across panorama generation tasks and models is left for future work.

Ethics statement

The development of advanced instruction-guided panoramic image generation models—which this study aims to evaluate and improve—carries significant ethical implications. While these technologies offer powerful capabilities for creative expression, they can also be misused to generate deceptive or harmful content (such as deepfakes, disinformation, or forged documents), thereby lowering the barrier for malicious actors to engage in harmful activities. By constructing more effective reward models, this study might inadvertently accelerate the development of these technological capabilities. We acknowledge this "dual-use" risk and have taken measures to mitigate it. Specifically, the PanoReward-Data dataset was constructed from public, non-sensitive benchmarks, with all personally identifiable information (PII) and sensitive content removed through a combination of automated and manual filtering. Our reward model, PanoReward, is trained to align with constructive, high-quality panorama generations defined by multidimensional evaluation criteria. Furthermore, all generated data and model outputs are released under the CC-BY-NC-SA 4.0 license, which explicitly prohibits commercial use, thereby reducing the risk of misuse such as deepfakes or other harmful applications. By making the dataset, models, and code publicly available, we aim to enhance transparency and support the research community in further investigating the safety, bias, and alignment characteristics of such models. Finally, we urge the community to adopt similar safeguards, including watermarking, provenance tracking, and rigorous training data filtering, when deploying or scaling instruction-guided panoramic image generation models.

Reproducibility statement

To ensure the reproducibility of our research, we provide the following details. All our reward models were trained on 8 NVIDIA H20 GPUs. Baseline models were evaluated using their official public codebases and recommended configurations. For proprietary models (such as GPT-5.5, the Gemini series, and GLM5.3), we accessed their APIs in September 2026; given the evolving nature of these models, we have archived specific outputs to ensure consistency. Our new dataset, PanoReward-Data, was constructed according to the detailed protocols described in the main text; both this dataset and our evaluation benchmark, PanoReward-Bench, will be released to the public. The complete codebase for training and evaluating PanoReward, along with the final weights of the fine-tuned Qwen3-VL, will be made available on GitHub and HuggingFace.

References

Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: Measuring multi-view consistency in generated images. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6034–6044. IEEE, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. Text2light: Zero-shot text-driven hdr panorama generation. *ACM Transactions on Graphics (TOG)*, 41(6):1–16, 2022.

Haoran Feng, Dizhe Zhang, Xiangtai Li, Bo Du, and Lu Qi. Dit360: High-fidelity panoramic image generation via hybrid training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23367–23377, 2026.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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 an-

swering. In *2023 ieee/cvf international conference on computer vision (iccv)*, pp. 20349–20360. IEEE, 2023.

Nikolai Kalischek, Michael Oechsle, Fabian Manhardt, Philipp Henzler, Konrad Schindler, and Federico Tombari. Cubediff: Repurposing diffusion-based image models for panorama generation. In *International Conference on Learning Representations*, volume 2025, pp. 81169–81181, 2025.

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. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Jeongeun Lee, Ryang Heo, and Dongha Lee. Personalized reward modeling for text-to-image generation. In *European Conference on Computer Vision*, pp. 451–470. Springer, 2026.

Taiming Lu, Tianmin Shu, Alan Yuille, Daniel Khashabi, and Jieneng Chen. Genex: Generating an explorable world. In *International Conference on Learning Representations*, volume 2025, pp. 52310–52335, 2025.

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, 2025.

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

Changgyoon Oh, Wonjune Cho, Yujeong Chae, Daehee Park, Lin Wang, and Kuk-Jin Yoon. Bips: Bi-modal indoor panorama synthesis via residual depth-aided adversarial learning. In *Computer Vision – ECCV 2022*, pp. 352–371, 2022.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pp. 2234–2242, 2016.

Xiancheng Sun, Mai Xu, Shengxi Li, Senmao Ma, Xin Deng, Lai Jiang, and Gang Shen. Spherical manifold guided diffusion model for panoramic image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5824–5834, 2025.

Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *Advances in neural information processing systems*, 36:51202–51233, 2023.

HunyuanWorld Team, Zhenwei Wang, Yuhao Liu, Junta Wu, Zixiao Gu, Haoyuan Wang, Xuhui Zuo, Tianyu Huang, Wenhuan Li, Sheng Zhang, et al. Hunyuanworld 1.0: Generating immersive, explorable, and interactive 3d worlds from words or pixels. *arXiv preprint arXiv:2507.21809*, 2025.

HY Team, Chenjie Cao, Xuhui Zuo, Zhenwei Wang, Yisu Zhang, Junta Wu, Zhenyang Liu, Yuning Gong, Yang Liu, Bo Yuan, et al. Hy-world 2.0: A multi-modal world model for reconstructing, generating, and simulating 3d worlds. *arXiv e-prints*, pp. arXiv–2604, 2026.

Yibin Wang, Yuhang Zang, Feng Han, Jiazi Bu, Yujie Zhou, Cheng Jin, and Jiaqi Wang. Unified personalized reward model for vision generation. *arXiv preprint arXiv:2602.02380*, 2026.

World Labs. Marble: A multimodal world model. https://www.worldlabs.ai/blog/marble-world-model, November 2025. World Labs Blog.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 54015–54029, 2024.

Keming Wu, Sicong Jiang, Max Ku, Ping Nie, Minghao Liu, and Wenhu Chen. Editreward: A human-aligned reward model for instruction-guided image editing. *arXiv preprint arXiv:2509.26346*, 2025.

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

Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2096–2105. IEEE, 2023b.

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. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11269–11277, 2026.

Shuai Yang, Jing Tan, Mengchen Zhang, Tong Wu, Gordon Wetzstein, Ziwei Liu, and Dahua Lin. Layerpano3d: Layered 3d panorama for hyper-immersive scene generation. In *Proceedings of the special interest group on computer graphics and interactive techniques conference conference papers*, pp. 1–10, 2025a.

Zhongqi Yang, Wenhang Ge, Yuqi Li, Jiaqi Chen, Haoyuan Li, Mengyin An, Fei Kang, Hua Xue, Baixin Xu, Yuyang Yin, et al. Matrix-3d: Omnidirectional explorable 3d world generation. *arXiv preprint arXiv:2508.08086*, 2025b.

Cheng Zhang, Qianyi Wu, Camilo Cruz Gambardella, Xiaoshui Huang, Dinh Phung, Wanli Ouyang, and Jianfei Cai. Taming stable diffusion for text to 360 panorama image generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6347–6357. IEEE, 2024a.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8018–8027. IEEE, 2024b.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. In *International Conference on Learning Representations*, volume 2026, pp. 134129–134150, 2026.

A Benchmark Construction Details

Scene selection and candidate generation. The benchmark contains 45 scenes selected to provide sufficient spatial context for panoramic evaluation. We retain coherent urban, street, indoor, natural, architectural, residential, and commercial environments. Close-ups and isolated fragments that do not provide enough surrounding context are removed. Each retained scene is paired with four candidate panoramas generated by different panorama systems (DiT360 (Feng et al., 2026), FLUX.2, HunyuanWorld-1 (Team et al., 2025), and Matrix3D (Yang et al., 2025b)) under the same scene description.

Annotation pool and ranking protocol. Ten annotators rank the four candidates in every cell, where rank 1 denotes the best candidate and rank 4 denotes the worst candidate. Thus, one annotator provides a complete 4-way ranking for each cell. Candidates are displayed in a randomly permuted anonymous order. The mapping from anonymous slots to source systems is recorded for later evaluation.

Hard and non-hard cells. A hard cell is a cell in which the human consensus does not provide a sufficiently clear winner. We mark a cell as hard when the rank-sum gap between the best and second-best candidate is smaller than 2. Every other cell is a non-hard cell. The main evaluation set therefore contains 35 G cells, 42 T cells, and 30 V cells. The excluded hard set contains 10 G cells, 3 T cells, and 15 V cells.

B Dataset and Benchmark Diversity

Scene and generator coverage of the training data. The source corpus covers 32,655 distinct scene descriptions. Each scene is realized by four panorama generation systems, and all six unordered system pairs are enumerated per scene, yielding 195,930 pairs in total; every system pair contributes exactly 32,655 pairs, so the training signal is fully crossed in the generator identity and no system pair is over-represented. Scene descriptions draw from a broad range of environments (urban, street, indoor, natural, architectural, residential, and commercial). Because the per-dimension scores are ordinal rankings within each scene, the G/V/T score distributions each cover the four levels evenly (25% per level), and the pairwise score margins span near-ties to clear wins: 10.1% of the pairs have equal overall scores, 23.8%, 21.1%, 16.6%, and 11.4% have margins of 1, 2, 3, and 4, and 16.9% have margins of 5 or more.

Description statistics. The scene descriptions vary widely in length and structure, with 125 words on average (median 51; 10th–90th percentile 32–371; range 3–781 words). 85% of them begin with a short title-style line followed by free-form elaboration of one to 32 paragraphs (median 2), so the model is trained under both terse and highly detailed text conditions. The benchmark captions are drawn from the same distribution (99 words on average, median 50, range 5–692), so the evaluation text conditions match the training ones in complexity.

Task-conditioned coverage. Table 6 reports the composition of the released training mixture. The three mixture components follow the 20%/30%/50% target of Section 4 (overall-profile, non-reversal counterfactual, and preference-reversal instances). Reversal samples are present under every strict profile, with shares between 55.3% and 68.9% among the counterfactual samples of each profile, so no profile can be solved by a fixed winner prior.

Scene and generator coverage of the benchmark. The 45 benchmark scenes cover the following environment categories: natural landscapes (20), architectural landmarks (9), urban and street scenes (8), interiors (5), residential scenes (2), and commercial scenes (1). Each scene is paired with the same four generation systems, giving 180 candidate panoramas, so every scene is judged under an identical generator crossing and scene difficulty is not confounded with generator identity.

Dimension discriminativeness. The identity of the best candidate differs across the three dimensions. Matrix3D wins the majority of G cells (27 of 45), FLUX wins most T cells (34 of 45), and DiT360 wins most V cells (23 of 45), while HunyuanWorld wins 11 G cells and 6 V cells. Thus no single system dominates all three dimensions, and the multi-dimensional decomposition is not redundant: a candidate that is best under one task instruction can be inferior under another. Consistent with this, every scene contains at least one *trade-off pair* (opposite G and T winners), with

Table 6: Composition of PanoReward-Data (943,074 training samples). Share of mixture is the fraction of all training samples. Reversal share is computed within the counterfactual samples of each profile; the overall profile has no counterfactual label and therefore no reversal share. The total reversal share is thus computed within the 754,460 counterfactual samples, which is 50.0% of all training samples and matches the 20%/30%/50% mixture target of Section 4.

Profile	Samples	Share of mixture	Reversal share
overall	188,614	20.0%	—
$G > V > T$	143,658	15.2%	68.9%
$G > T > V$	142,970	15.2%	68.3%
$V > G > T$	123,357	13.1%	61.2%
$V > T > G$	120,996	12.8%	61.2%
$T > G > V$	112,555	11.9%	55.3%
$T > V > G$	110,924	11.8%	57.0%
Total	943,074	100.0%	62.5%

1 to 6 such pairs per scene (180 in total), which supports the instruction-following evaluation of Section **??**.

Difficulty spectrum. The consensus margin (top-1 vs. top-2 rank sum) has a median of 4 on G, 5 on T, and 4 on V, spanning near-ties to clear majorities. Cells whose margin falls below the reliability threshold are kept as the hard set rather than discarded (Appendix A), so the benchmark retains an explicit difficulty axis.

C Evaluation Protocols and Metrics

Pairwise evaluation. Each non-hard 4-way ranking is expanded into three comparisons between the human winner and each of the other candidates. This produces 105 G comparisons, 126 T comparisons, and 90 V comparisons. The total is 321 comparisons. A model receives one point for a correct winner, zero points for an incorrect winner, and 0.5 points for a score tie. Pairwise accuracy is the micro-averaged score across comparisons.

Full 4-way evaluation. For each cell, a model assigns one score to each candidate. The candidates are sorted by these scores to obtain a predicted 4-way ranking. top-1 accuracy measures whether the predicted first candidate matches the human winner. Kendall's τ measures the agreement between the full predicted and human rankings. The random top-1 baseline is 0.25.

Human rankings are computed with leave-one-annotator-out agreement. For each annotator, we construct the Borda consensus from the other retained annotators in the same cell. We then compare the annotator ranking with this consensus and average over all annotators and cells.

Table 7 reports the full 4-way ranking results. PanoReward reaches 0.570 top-1 accuracy and 0.533 Kendall's τ overall, with the best top-1 on G (0.429) and T (0.714). The variant without task conditioning falls to 0.383 top-1 overall.

Table 7: Full 4-way ranking results on the non-hard cells. Each entry reports top-1 accuracy and Kendall's τ. Random top-1 accuracy is 0.25.

Method	G	T	V	All
Random	0.250 / 0.000	0.250 / 0.000	0.252 / 0.000	0.250 / 0.000
Qwen3-VL-8B	0.200 / −0.090	0.214 / 0.252	0.067 / −0.011	0.168 / 0.050
Qwen3.5-9B	0.171 / −0.346	0.333 / 0.356	0.233 / 0.066	0.246 / 0.025
PanoReward w/o task-profile supervision	0.000 / −0.394	0.619 / 0.492	0.500 / 0.287	0.383 / 0.128
PanoReward (ours)	**0.429 / 0.503**	**0.714 / 0.627**	0.533 / 0.469	**0.570 / 0.533**

D Implementation Details

PanoReward uses Qwen3-VL-8B-Instruct (Bai et al., 2025) as the backbone. The three dimension heads read the <|Reward|> representation and predict (m_k, ℓ_k) for each dimension. The fusion branch is a two-layer MLP with hidden size 32. Its output layer is zero initialized. We train with the uncertainty-aware Bradley-Terry objective from Section 3 for one epoch with learning rate 2×10^{-6}, 5% warm-up, per-device batch size 1, gradient accumulation 4, and bf16 precision on 8 GPUs. Images use a fixed pixel budget of $524{,}288$ or 1024×512.

For reward inference, each panorama is scored independently. The score therefore does not depend on candidate order. For reinforcement-learning post-training, PanoReward is frozen and one DiT360 policy is trained for each reward condition.

E Input Templates

Each dimension head receives its fixed quality rubric and the scene description. The task instruction is not provided to the three dimension heads. The fusion branch receives the three rubric summaries together with the task instruction.

> **G (3D consistency).** "You are evaluating a generated 360-degree panorama against the supplied text description. Assign a quality score from 1 to 4, where 4 is best. Focus on 3D consistency and continuity in the middle viewing region: stable perspective, coherent structures across neighboring views, natural proportions, and absence of seams, stretching, tearing, or duplicated content. Also check that the main scene agrees with the text description."
>
> **V (polar visual quality).** "You are evaluating a generated 360-degree panorama. Assign a quality score from 1 to 4, where 4 is best. Focus on the top and bottom pole regions. Prefer natural, clean content with little noise, blur, stretching, swirling distortion, broken structure, or unusual color blocks. Judge the visible panorama itself. There is no reference image."
>
> **T (image-text consistency).** "You are evaluating a generated 360-degree panorama against the supplied text description. Assign a quality score from 1 to 4, where 4 is best. Focus on image-text consistency and appearance quality in the middle viewing region. The described objects, layout, lighting, and atmosphere should be present, while textures should be coherent, detailed, and free of obvious generation artifacts or discontinuities."

The fusion branch is trained with ten paraphrases for each of the six strict priority profiles and two held-out paraphrases for validation. The overall profile uses ten training paraphrases and two held-out paraphrases that express equal importance. The numerical weights used by the training procedure do not appear in the natural-language instructions. Representative instructions are listed below.

> $G > V > T$: "Prioritize 3D consistency, then pole-region visual quality, then text adherence."
>
> $V > T > G$: "This panorama will be viewed at the zenith and nadir, so pole-region quality is most important, text adherence is next, and geometry is last."
>
> $T > V > G$: "If the described objects or layout are missing, that is the worst failure. Prioritize text fit, then pole quality, then geometry."
>
> Overall: "Judge overall panorama quality. The three aspects should be weighted uniformly."

F Extended Results

Aggregation protocol grid. Table 8 reports the complete aggregation comparison. Uniform averaging gives an instruction-free reference. Fixed weighting applies the priority-position template $(8, 4, 1)$. The head protocol uses the score from the dimension corresponding to the evaluated criterion. Profile uses the learned task-profile-aware fusion.

Table 8: Complete aggregation comparison on the main evaluation set. Values are pairwise accuracy.

Model	Aggregation	G	T	V	All
Without task-profile supervision	uniform	0.324	0.770	0.700	0.604
Without task-profile supervision	fixed $(3, 1, 1)$	0.305	0.770	0.689	0.595
Without task-profile supervision	fixed $(8, 4, 1)$	0.324	0.770	0.689	0.601
Without task-profile supervision	head	0.305	0.770	0.667	0.589
PanoReward	uniform	0.381	0.865	0.700	0.660
PanoReward	fixed $(8, 4, 1)$	0.686	0.722	0.722	0.710
PanoReward	head	0.714	0.786	0.778	0.760
PanoReward	learned fusion (ours)	**0.705**	**0.857**	**0.778**	**0.785**

Instruction generalization. Table 9 evaluates whether the fusion branch responds to the meaning of the instruction rather than memorizing a fixed phrase. The learned weights vary little across training paraphrases. Held-out paraphrases preserve the intended highest-priority dimension in all evaluated profiles.

Table 9: Instruction generalization on training and held-out paraphrases. $\alpha = (\alpha_G, \alpha_V, \alpha_T)$ follows the dimension order (G, V, T). Spread is the maximum absolute deviation across paraphrases.

Profile	Phrasings	Accuracy	α (mean)	Spread
$G > V > T$	10 train	0.705 ± 0.000	(0.735, 0.248, 0.017)	0.007
	2 held-out	0.710	(0.739, 0.243, 0.018)	0.004
$V > T > G$	10 train	0.779 ± 0.003	(0.015, 0.784, 0.201)	0.014
	2 held-out	0.789	(0.117, 0.768, 0.115)	0.090
$T > V > G$	10 train	0.857 ± 0.000	(0.022, 0.276, 0.701)	0.016
	2 held-out	0.770	(0.257, 0.088, 0.655)	0.064
overall	10 train	0.651 ± 0.000	(0.134, 0.437, 0.429)	0.016
	2 held-out	0.657	(0.174, 0.467, 0.358)	0.071

Disentanglement of the head. Without task-conditioned supervision, the three dimension heads become highly similar on benchmark images, with pairwise Spearman correlations between 0.96 and 0.99. The G head also reaches only 0.324 agreement with the G ground truth. With task-conditioned supervision, the corresponding head correlations are -0.72, -0.60, and -0.02, and the three heads recover distinct dimension-specific behavior.

G Analysis and Limitations

We first report diagnostic analyses of model and baseline behavior on the benchmark, then state the limitations of the current model.

Diagnostic analysis: ambiguous cases. The hard cells are intentionally excluded from the headline benchmark results because their human consensus does not provide a sufficiently clear winner. On the 45 hard V comparisons (15 hard V cells, three comparisons per cell), pairwise agreement ranges from 0.51 to 0.67 across the evaluated systems, compared with 0.78 for PanoReward on the main V set. We report these cells separately to avoid mixing annotation ambiguity with model error.

Limitation: reward hacking risk. As with any learned reward model, downstream optimization can exploit residual evaluator biases. We therefore recommend evaluating optimized generators with both PanoReward and independent human judgments. The three dimension scores should also be monitored jointly rather than optimizing only the scalar reward.

Limitation: sensitivity to compressed instructions. On held-out instruction phrasings, the T-priority accuracy drops from 0.857 to 0.770 (Table 9); the top-priority dimension is always identified, but secondary weights can shift under very compact paraphrases.

Limitation: benchmark coverage. PanoReward-Bench covers 45 scenes and four generation systems; broader scene types, more generators, and larger annotation pools would further tighten the conclusions.

Table 10: Presentation-order bias of VLM judges. The gap is the accuracy difference between the two presentation orders. Swap consistency is the fraction of pairs whose verdict is unchanged after the order is swapped. PanoReward scores each panorama independently, so its verdict is invariant to presentation order (gap 0 and swap consistency 1 by construction). Its accuracies are computed on the same probe judgments as the judges and on the same per-criterion scoring subsets, which differ slightly from the main evaluation set.

Judge	Criterion	Winner first	Winner second	Gap / Swap consistency
Qwen3-VL-8B	G	0.356	0.239	0.117 / 0.839
	V	0.656	0.600	0.056 / 0.878
	T	0.894	0.774	0.120 / 0.862
Qwen3.5-9B	G	0.333	0.194	0.139 / 0.839
	V	0.711	0.539	0.172 / 0.817
	T	0.931	0.811	0.120 / 0.880
PanoReward (ours)	G	0.733	0.733	0.000 / 1.000
	V	0.778	0.778	0.000 / 1.000
	T	0.871	0.871	0.000 / 1.000

H Presentation-Order Bias

Human ranking is protected against presentation-order effects by anonymous per-case permutation. PanoReward scores each panorama independently and is therefore order-free at inference. For general VLM judges, we perform a dedicated order-swap probe in which the same pair is judged in both presentation orders. There are 642 judgments for each judge across the 321 main pairs. We report accuracy when the ground-truth winner is shown first, accuracy when the winner is shown second, and swap consistency, which is the fraction of pairs whose verdict remains unchanged after swapping the two candidates. Table 10 reports the results. The PanoReward rows confirm that order-free scoring removes the bias entirely.

Note that the accuracies in Table 10 are not directly comparable with Table 2. In the swap probe, every pair is judged under all six criteria, and a criterion is scored on the subset of probe comparisons that carry a reference label for that criterion. Because the same pair of systems can be compared in more than one dimension cell of the same scene, these subsets are not identical to the 105 G, 126 T, and 90 V comparisons of the main evaluation, and the accuracies therefore differ slightly (e.g., 0.733 vs. 0.705 on G). The two tables serve different purposes: Table 2 measures preference agreement, whereas Table 10 measures presentation-order robustness on the probe set.

Furthermore, general VLM judges show limited instruction sensitivity on the trade-off set. The three preference-conditioned scores are identical to the dimension scores in 180 of 180 pairs for Qwen3-VL-8B and 177 of 180 pairs for Qwen3.5-9B. Switching the instruction changes the verdict in only 1 of 45 and 8 of 45 scenes, respectively. This result motivates explicit task-conditioned training rather than relying on prompting alone.

I AI Assistance

We used AI to polish the grammar of this paper. AI was not used to assist with the core ideas or the experimental design.

J Qualitative Examples

The following figures show qualitative examples of PanoReward.

(a) DiT360



Profile	G > V > T	V > T > G	T > V > G
R (rank)	+0.62 (3)	+0.32 (2.5)	−1.86 (3)

(b) FLUX.2



Profile	G > V > T	V > T > G	T > V > G
R (rank)	−1.09 (4)	**+3.07 (1)**	**+2.06 (1)**

(c) HunyuanWorld 1.0



Profile	G > V > T	V > T > G	T > V > G
R (rank)	**+2.33 (1)**	−1.16 (2.5)	−3.48 (4)

(d) Matrix3D



Profile	G > V > T	V > T > G	T > V > G
R (rank)	+0.57 (2)	+2.75 (4)	−0.90 (2)

Prompt excerpt This is a sun-drenched, wide-angle photograph capturing the majestic facade and western side of the Cathedral of San Giorgio (Duomo di Montepulciano) in Tuscany, Italy. The building stands as a powerful example of Romanesque architecture, its warm, honey-toned stone walls glowing under a brilliant azure sky dotted with fluffy white cumulus clouds.

R ↑: task-conditioned reward. Parentheses: reference rank ↓. Bold: highest R within each profile.
G: 3D geometric consistency V: polar visual quality T: image–text consistency
G/T reference: human Borda ranking. V reference: human + NIQE. Rank 1 is best.

(a) DiT360



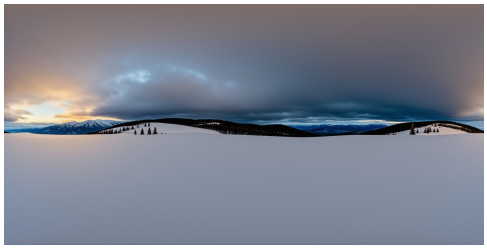
Profile	G > V > T	V > T > G	T > V > G
R (rank)	+0.30 (3)	−0.54 (2.5)	−0.34 (2)

(b) FLUX.2



Profile	G > V > T	V > T > G	T > V > G
R (rank)	−1.47 (4)	**+2.79 (1)**	**+3.11 (1)**

(c) HunyuanWorld 1.0



Profile	G > V > T	V > T > G	T > V > G
R (rank)	**+1.56 (1)**	−3.22 (4)	−0.99 (4)

(d) Matrix3D



Profile	G > V > T	V > T > G	T > V > G
R (rank)	+0.74 (2)	−0.37 (2.5)	−2.25 (3)

Prompt excerpt This is a serene and atmospheric wide-angle landscape photograph capturing a snow-covered rural valley at either dawn or dusk, with majestic, rolling hills in the background under a dynamic, cloud-filled sky.

R ↑: task-conditioned reward. Parentheses: reference rank ↓. Bold: highest R within each profile.
G: 3D geometric consistency V: polar visual quality T: image–text consistency
G/T reference: human Borda ranking. V reference: human + NIQE. Rank 1 is best.

(a) DiT360



Profile	G > V > T	V > T > G	T > V > G
R (rank)	−1.37 (3)	−1.07 (2)	+2.04 (2)

(b) FLUX.2



Profile	G > V > T	V > T > G	T > V > G
R (rank)	−1.77 (4)	+0.03 (3)	**+2.41 (1)**

(c) HunyuanWorld 1.0



Profile	G > V > T	V > T > G	T > V > G
R (rank)	**+0.16 (2)**	−0.23 (4)	−0.93 (4)

(d) Matrix3D



Profile	G > V > T	V > T > G	T > V > G
R (rank)	−0.95 (1)	**+0.52 (1)**	+2.07 (3)

Prompt excerpt This panoramic photograph captures a breathtaking expanse of terraced rice paddies cascading down a steep, verdant mountainside, likely in a region such as the Himalayas, Southeast Asia, or the Andes — areas renowned for this agricultural artistry.

R ↑: task-conditioned reward. Parentheses: reference rank ↓. Bold: highest R within each profile.
G: 3D geometric consistency V: polar visual quality T: image–text consistency
G/T reference: human Borda ranking. V reference: human + NIQE. Rank 1 is best.

(a) DiT360



Profile	G > V > T	V > T > G	T > V > G
R (rank)	−1.24 (3)	+0.22 (2)	+1.11 (3)

(b) FLUX.2



Profile	G > V > T	V > T > G	T > V > G
R (rank)	−1.34 (4)	+0.46 (3)	**+1.26 (1)**

(c) HunyuanWorld 1.0



Profile	G > V > T	V > T > G	T > V > G
R (rank)	−0.16 (1)	−3.36 (4)	−0.64 (4)

(d) Matrix3D



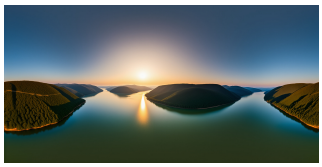
Profile	G > V > T	V > T > G	T > V > G
R (rank)	**+0.02 (2)**	**+2.45 (1)**	−0.29 (2)

Prompt excerpt This is a vibrant, sun-drenched street corner in a historic European city — most likely Madrid, Spain — characterized by its rich architectural heritage, bustling pedestrian life, and warm, golden-hour lighting. The composition captures the dynamic interplay between architecture, human activity, and natural elements under a clear blue sky.

R ↑: task-conditioned reward. Parentheses: reference rank ↓. Bold: highest R within each profile.
G: 3D geometric consistency V: polar visual quality T: image–text consistency
G/T reference: human Borda ranking. V reference: human + NIQE. Rank 1 is best.

Low PanoReward scores

R ↑ | reference rank ↓

3D geometric consistency G > V > T	Polar visual quality V > T > G	Image–text consistency T > V > G
FLUX.2	**HunyuanWorld 1.0**	**HunyuanWorld 1.0**
		
R = −2.17 Ref. rank 4 / 4	R = −3.89 Ref. rank 4 / 4	R = −4.46 Ref. rank 4 / 4
Black-and-white arcade	Rhine River at dawn	Winter lake in the Rockies
FLUX.2	**HunyuanWorld 1.0**	**HunyuanWorld 1.0**
		
R = −1.81 Ref. rank 4 / 4	R = −3.36 Ref. rank 4 / 4	R = −4.41 Ref. rank 4 / 4
Mediterranean coastal town	European city street	Historic European courtyard

R: reward under the column's priority instruction; larger is better. Examples use non-hard reference cells.
G/T reference: human Borda ranking. V reference: human + NIQE. Scene captions are provided with the figures.

High PanoReward scores

R ↑ | reference rank ↓

3D geometric consistency G > V > T	Polar visual quality V > T > G	Image–text consistency T > V > G
HunyuanWorld 1.0	**FLUX.2**	**FLUX.2**
		
R = +2.33 Ref. rank 1 / 4	R = +3.49 Ref. rank 1 / 4	R = +3.11 Ref. rank 1 / 4
Cathedral in Tuscany	Autumnal river valley	Snow-covered valley
Matrix3D	**FLUX.2**	**FLUX.2**
		
R = +2.03 Ref. rank 1 / 4	R = +3.34 Ref. rank 1 / 4	R = +2.95 Ref. rank 1 / 4
Sunlit river landscape	Stone archway and courtyard	Lighthouse keepers' residence

R: reward under the column's priority instruction; larger is better. Examples use non-hard reference cells.
G/T reference: human Borda ranking. V reference: human + NIQE. Scene captions are provided with the figures.

PanoReward ranking failures

Model preference vs. reference preference

3D geometric consistency G > V > T	Polar visual quality V > T > G	Image–text consistency T > V > G
Model prefers: HunyuanWorld 1.0	**Model prefers: FLUX.2**	**Model prefers: FLUX.2**
		
R = +1.93 Ref. rank 3 / 4	**R = +2.44** Ref. rank 2.5 / 4	**R = +3.38** Ref. rank 3 / 4
Reference prefers: Matrix3D	**Reference prefers: HunyuanWorld 1.0**	**Reference prefers: DiT360**
		
R = −0.58 Ref. rank 1 / 4	**R = −2.80** Ref. rank 1 / 4	**R = −1.79** Ref. rank 1 / 4
The Church of St. John's	Urban building facade	Autumnal hills and dramatic sky
Reward reversal: ΔR = +2.50	**Reward reversal: ΔR = +5.24**	**Reward reversal: ΔR = +5.17**

ΔR = R(model-preferred) − R(reference-preferred). Positive margins show disagreement with the reference.
All cells are non-hard. G/T: human Borda reference. V: human + NIQE reference. Rank 1 is best.