PosterReward: Unlocking Accurate Evaluation for High-Quality Graphic Design Generation

Supplementary Material

This is supplementary material for ***PosterReward: Unlocking Accurate Evaluation for High-Quality Graphic Design Generation.***

We present the following materials in this supplementary material:

- **Sec.1** Qualitative comparison of Best-of-8 selection performance against HPSv3, UnifiedReward, and MLLMs, highlighting robustness to scoring failures and position bias.
- **Sec.2** Visual and quantitative comparison of SD3.5-Medium fine-tuned via Diffusion-NFT using PosterReward versus baselines like HPSv3, UnifiedReward, and PaddleOCR.
- **Sec.3** Analysis of the dataset construction pipeline, including multi-model verification, position bias mitigation strategies, and quality assessment via Kendall's Coefficient of Concordance.
- **Sec.4** Investigation into the impact of dataset scale, filtering stringency, and the integration of general aesthetic data on downstream model performance.
- **Sec.5** Detailed training hyperparameters within the MS-Swift framework and construction specifications for the PosterRewardBench Advanced and Basic subsets.
- **Sec.6** Discussion on computational overhead and MLLM evaluation constraints, alongside future directions for efficiency and scalability.

1. Test-Time-Scaling with PosterReward via Best-of-8 Selection

We investigate the feasibility of using PosterReward for poster selection. Based on the poster data generated by various models in the PosterBench benchmark, we selected several samples and apply Test-Time-Scaling via Best-of-8 (Bo8) selection. The results are presented in Figure 1.

As illustrated in the figure, PosterReward demonstrates a more accurate assessment of the poster content. Compared to HPSv3[6], PosterReward can more precisely evaluate text accuracy and layout requirements, showing a significant advantage in graphic design scenarios. UnifiedReward[9] is unsuitable for Bo8 selection, as it frequently produces identical scores for different candidates, making it difficult to distinguish the best one. We also attempted Bo8 selection using Multi-modal Large Language Models (MLLMs), which involved providing multiple images as input and prompting the model to identify the index of the optimal one. Our experiments indicate that the MLLMs struggle to make accurate choices among several visually similar images. Furthermore, we observe that even state-of-the-art proprietary MLLMs exhibit position bias[5, 7] when selecting among multiple similar images. Taking Gemini-2.5-Pro[2] as an example, when the first image in a sequence is of sufficient quality, the model demonstrates a tendency to select it, with its Chain-of-Thought (CoT) predominantly focusing on justifying the first image as the optimal candidate. The prompt for using MLLM for Bo8 selection is shown in Figure 8.

2. Reinforcement Learning with PosterReward via Diffusion-NFT

With the increasing prevalence of post-training reinforcement learning (RL) in image generation, numerous evaluation models have been adopted as reward models to furnish optimization signals. Consequently, we evaluate the efficacy of RL utilizing PosterReward. We select DiffusionNFT[11] as our reinforcement learning paradigm. Unlike conventional Policy Gradient methods that discretize the reverse sampling process, DiffusionNFT optimizes the diffusion model directly on the forward process via the flow matching objective. It contrasts positive and negative generations to define an implicit policy improvement direction, naturally incorporating reinforcement signals into a supervised learning objective:

$$\mathcal{L}(\theta) = \mathbb{E}_{\boldsymbol{c},\pi^{\text{old}},t}\Big[r\|\boldsymbol{v}_\theta^+(\boldsymbol{x}_t,\boldsymbol{c},t) - \boldsymbol{v}\|_2^2 + (1-r)\|\boldsymbol{v}_\theta^-(\boldsymbol{x}_t,\boldsymbol{c},t) - \boldsymbol{v}\|_2^2\Big] \tag{1}$$

where $\boldsymbol{v}_\theta^+$ and $\boldsymbol{v}_\theta^-$ denote the implicit positive and negative velocities, respectively. This formulation enables efficient optimization compatible with arbitrary black-box solvers without requiring likelihood estimation.

We benchmark against reward models including HPSv3[6] and UnifiedReward-Qwen-7B. Both output scalar scores, representing state-of-the-art discriminative and generative reward paradigms, respectively. Furthermore, given our specific focus on poster and graphic design, we incorporate PaddleOCR as a reward model; it provides signals by evaluating the Levenshtein edit distance between the recognized text and the ground-truth text required for rendering. We employ Stable Diffusion 3.5 Medium (SD3.5-M) as the base model. The fine-tuning process utilizes 8 NVIDIA A100 GPUs, with an additional 4 A100 GPUs dedicated to reward model deployment. All training hyperparameters adhere to the original DiffusionNFT

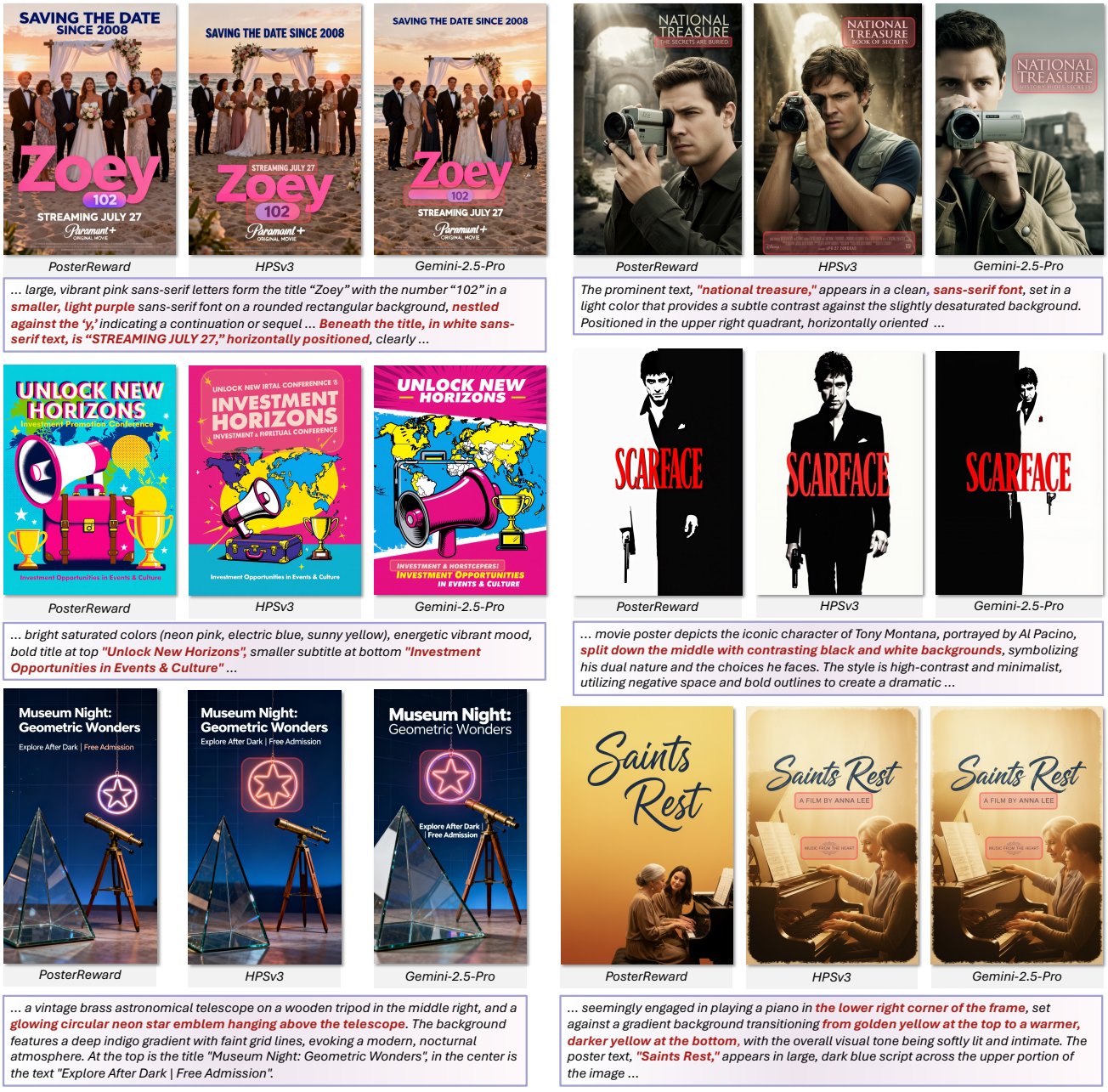

Figure 1. **Test-Time-Scaling results using Best-of-8 (Bo8) Selection.** Comparison of poster selection performance between PosterReward, HPSv3, and Gemini-2.5-Pro. PosterReward demonstrates superior accuracy in assessing text and layout quality, avoiding the scoring failures and position biases observed in other models.

configuration. We select the model fine-tuned for 350 steps for evaluation. For the original SD3.5-M model, we followed the setting of CFG=4.5, while for the fine-tuned model, we used CFG=1.0. The seed used during generation is always fixed at 0.

For completeness, we additionally present the Flux.1-dev visual comparison in Figure 4. We present a comprehensive evaluation of SD3.5-M optimized by various reward models, with visual comparisons detailed in Figure 2 and Figure 3, and quantitative user study results summarized in Figure 5. Regarding **text accuracy**, although HPSv3[6] and UnifiedReward[9] effectively improve the rendering of key headlines, they inadvertently introduce a substantial amount of illegible small text—resembling

*This vibrant, cinematic poster for "**Standing Up, Falling Down**" features two male subjects framed against a bright yellow background, divided by a microphone stand at the center. On the left, an older man with a beard wears a dark fedora and a purple scarf; on the right, a younger man with a slight smile looks towards the viewer. The poster's energetic style uses a warm, optimistic color palette to suggest a comedic or feel-good narrative. The title, "STANDING UP FALLING DOWN," is prominently **displayed in large, bold, white uppercase sans-serif font, positioned at the top and horizontally centered**, conveying the film's central theme of resilience and overcoming obstacles. The stark white text contrasts sharply with the yellow background, ensuring high visibility and immediate recognition. The layout strategically **places the two subjects on either side of the central microphone**, implying their connection within the story, likely related to comedy or performance. The large title at the top dominates the visual hierarchy, directing the viewer's attention to the film's identity before drawing them into the character portraits below. The overall composition creates a sense of balance while hinting at the dynamic between the two figures.*

*The poster for "Night Hunter" is a dark, atmospheric composition **featuring four figures** against a background of fractured ice and fog, suggesting a chilling mystery. The poster adopts a cinematic, dramatic style, utilizing a cool, desaturated color palette dominated by deep greens, blues, and grays. The title, **"NIGHT HUNTER,"** is centrally positioned in two lines of text, the top word "NIGHT" in a smaller, lighter teal sans-serif font and the bottom word "HUNTER" in a much larger, bold white sans-serif font. Both words are presented horizontally and contrast starkly with the dark, misty background, creating a strong visual anchor. The top of the poster is dominated by the **brooding face of a man in profile**, partially obscured by swirling darkness, his gaze directed towards the right. Below him and to the left are two other male figures, rendered in profile and shown from the chest up, their expressions serious. A woman's face is positioned between these figures and the central man, her striking blue eyes looking directly forward. At the bottom of the poster, a solitary figure stands beside a stack of logs in a desolate, foggy landscape. The layout is layered, with the figures positioned in the upper and middle sections, progressively receding into the background towards the top, while the landscape grounds the scene at the bottom, all unified by the encompassing, icy and foggy texture that envelops the entire frame.*

*This compelling poster for *Fight Club* features a double exposure illustration of two male figures, one in profile with closed eyes and another facing mostly forward, overlaid with a sense of motion blur, conveying a feeling of internal conflict and a fragmented identity, **all rendered in a bold, graphic style with distinct outlines and color blocking. The overall aesthetic is a blend of illustration and cinematic abstraction, utilizing a limited color palette of muted greens, blues, and deep reds to create a moody and unsettling atmosphere.** The text "FIGHT CLUB" is centrally positioned near the bottom, written in a clean, sans-serif font that appears to glow with a bright pinkish-red neon effect, suggesting the clandestine and vibrant nature of the underground activity. The letters are slightly distorted and angled, enhancing the sense of motion present in the illustration, and it is oriented horizontally. The layout is vertically dominant, with the figures occupying the upper and central portions of the frame, and the title serving as a grounding element at the bottom. The overlapping figures and the motion blur create a dynamic visual tension that draws the viewer's eye upwards, while the glowing title acts as a focal point, reinforcing the central theme of the film.*

Figure 2. **Visual comparison of SD3.5-Medium fine-tuned with various reward models (Part 1).** From left to right, the columns display the outputs of the original SD3.5-Medium, followed by models fine-tuned using PosterReward, HPSv3, UnifiedReward, PaddleOCR, and the combined UnifiedReward + PaddleOCR. The corresponding prompts are enclosed in the purple boxes at the bottom.

"credit blocks"—which severely compromises the overall visual coherence. In contrast, PosterReward significantly elevates text rendering quality while effectively preventing the generation of redundant content. Furthermore, while incorporating PaddleOCR[3] as a fine-tuning signal yields higher character recognition accuracy, it compromises textual aesthetics (e.g., font style and integration) and exhibits a lower success rate when handling complex, multi-text ren-

A woman with dark hair, styled messily and framing her face, is depicted in a close-up, looking directly at the viewer with a serious expression. Her left hand is raised, touching the back of her head amidst her hair, while her right arm is visible at the bottom left of the frame. ***A bandage is placed across the bridge of her nose****. The backdrop is a* ***muted yellow wall with a grid pattern of light gray tiles****, suggesting an interior setting, possibly a bathroom or institutional space, contributing to the poster's raw and slightly unsettling visual tone. The overall style is cinematic and stark, with a raw, vérité feel due to the close-up perspective and the visible bandage. Prominently positioned at the top and center of the image is the title* ***"SISTER MIDNIGHT"****. The communicative intent of this text is clearly to identify the film. The text is rendered in a clean, sans-serif font, appearing in a bright white against the slightly desaturated yellow and the woman's dark hair. The letters have a crisp edge and solid texture, contrasting sharply with the grainy, slightly out-of-focus background. Both lines of text are horizontally aligned. The layout strategically places the text above the woman's head, drawing immediate attention to the title while the woman's gaze and the visual detail of the bandage create a compelling focal point at the center of the frame, hinting at vulnerability or a challenging narrative.*

The poster features a stark visual of a woman's face, illuminated by a vibrant blue light that wraps around her form, positioned diagonally across a dark background, creating a sense of dynamism and intrigue, giving the poster a cinematic and slightly mysterious style. ***The dominant text in the poster is the word "CIRCUIT," which serves as the title, presented in bold, sans-serif block letters with a metallic, slightly weathered texture, evoking a sense of urban grittiness or digital decay****. These letters are arranged horizontally across the lower central portion of the poster and are angled slightly upwards, mirroring the diagonal line of the figure, further integrating the text with the overall visual composition. The layout employs a strong diagonal line that draws the viewer's eye from the top left, across the illuminated figure, and down to the prominent title at the bottom, creating a sense of movement and emphasizing the central subject and the film's title in a cohesive and impactful design.*

A neumorphic poster for culture-tech journal recommendations, featuring a leather-bound book (left), sleek tablet (center), globe (right), and stylus (below) on a soft blue-gray gradient surface. At the top is the title ***"Top Journal Picks: Culture & Tech"****; in the center is the text* ***"Where Heritage Inspires Tomorrow"****.*

Figure 3. **Visual comparison of SD3.5-Medium fine-tuned with various reward models (Part 2).** Consistent with the previous figure, the columns represent the original SD3.5-Medium, PosterReward, HPSv3, UnifiedReward, PaddleOCR, and the joint UnifiedReward + PaddleOCR, respectively. Text prompts are provided in the bottom purple boxes.

dering scenarios. In terms of **aesthetics and composition**, PosterReward provides superior and more precise reward signals, whereas HPSv3 underperforms in both dimensions. While UnifiedReward exhibits a relatively high standard of aesthetics and composition, its scores are diminished by text accuracy defects and a noticeable reduction in image realism. Additionally, the dual-reward model combining PaddleOCR and UnifiedReward marginally sacrifices aesthetic and compositional quality to prioritize text correctness. Consequently, in terms of **overall preference**, the model fine-tuned with PosterReward successfully outperforms all competitors, demonstrating its capability to provide high-quality, holistic reward signals for poster generation tasks.

Figure 4. **Visual comparison of Flux.1-dev fine-tuned with various reward models.** From left to right, the columns display the outputs of the original Flux.1-dev, followed by models fine-tuned using **PosterReward**, **HPSv3**, **UnifiedReward** and **PickScore**. The corresponding prompts are enclosed in the purple boxes on the right.

3. Details and Quality Analysis of the AI Preference Dataset

In the preference data pipeline, we employ four Multimodal Large Language Models (MLLMs): GLM-4.5v[8], GPT-5, Gemini-2.5-Flash-lite, and Gemini-2.5-Pro. The "Thinking" mode is activated for all models; specifically, for the Gemini-2.5-Flash-lite and Gemini-2.5-Pro, the thinking budget is set to 4096. Regarding sampling parameters, the temperature is set to 1.0 for GPT-5, while a value of 0.7 is maintained for the remaining models. The specific prompts utilized for ranking and preference judgment are presented in Figure 13, Figure 14 and Figure 15.

We employ Gemini-2.5-Pro to annotate data for both single-image analysis and the collection of paired-image preference Chain-of-Thought (CoT) sequences required for fine-tuning. The detailed prompts utilized for these tasks are presented in Figure 9 and Figure 10. Specifically, during the acquisition of preference CoTs, we explicitly provide the ground-truth preference labels to the model, instructing

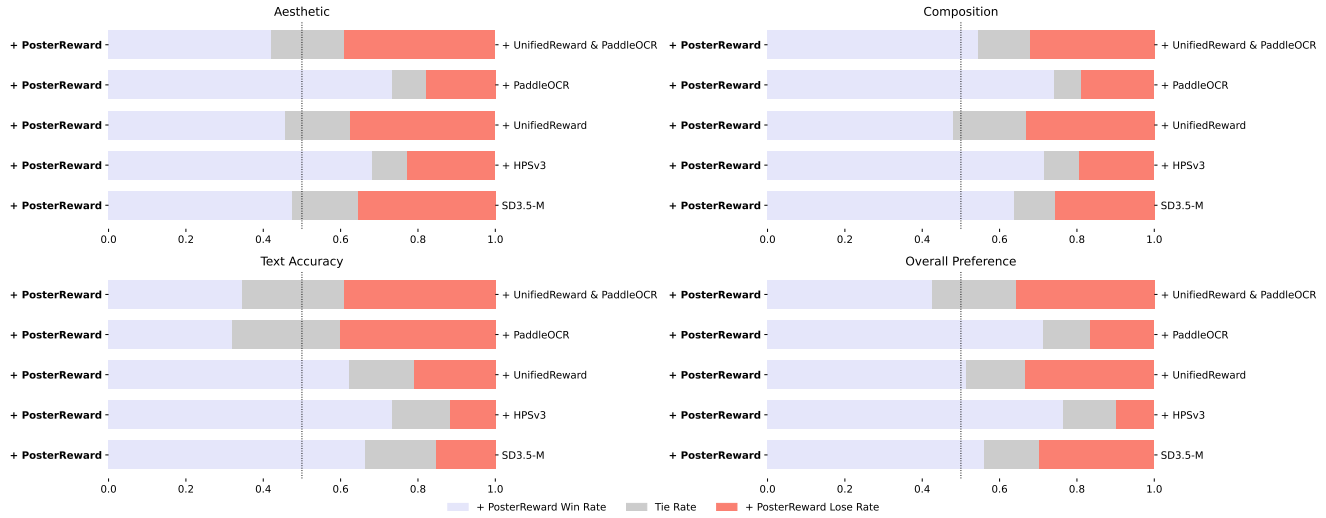

Figure 5. **User study results comparing PosterReward with varying baselines.** The evaluation covers four dimensions: Aesthetic, Composition, Text Accuracy, and Overall Preference. The bars represent the win (purple), tie (gray), and loss (red) rates of the model fine-tuned with PosterReward against SD3.5-M, HPSv3, UnifiedReward, and PaddleOCR[3]. PosterReward demonstrates a consistent preference advantage across most metrics.

it to articulate the reasoning behind the selection.

To evaluate the visual quality and semantic alignment of the generated movie posters, we employed a multi-round ranking strategy utilizing the Gemini-2.5-Flash-lite multimodal large language model. Given the stochastic nature of Large Language Models (LLMs), a single inference may not accurately reflect the model's stable preference. Therefore, for each creative brief, we presented a set of $m = 6$ candidate images to the model and instructed it to rank them from best to worst based on specific criteria (e.g., prompt adherence, aesthetic quality, and text rendering). This process was repeated $k = 6$ times independently to obtain a distribution of rankings.

To quantify the reliability and consistency of the model's rankings across these k iterations, we calculated Kendall's Coefficient of Concordance (Kendall's W)[4]. Kendall's W is a non-parametric statistic that assesses agreement among raters—in this context, the distinct inference rounds act as pseudo-raters.

Let m be the number of images (items) and k be the number of ranking rounds (raters). Let $r_{i,j}$ denote the rank assigned to the i-th image in the j-th round, where $1 \leq r_{i,j} \leq m$. The sum of ranks, R_i, given to the i-th image is calculated as:

$$R_i = \sum_{j=1}^{k} r_{i,j} \tag{2}$$

The mean of the sum of ranks, $\bar{R}$, is given by:

$$\bar{R} = \frac{1}{m} \sum_{i=1}^{m} R_i = \frac{k(m+1)}{2} \tag{3}$$

Kendall's W is defined as the ratio of the observed sum of squared deviations (S) to the maximum possible sum of squared deviations. First, we calculate S:

$$S = \sum_{i=1}^{m} (R_i - \bar{R})^2 \tag{4}$$

Assuming no ties in the rankings, Kendall's W is computed as:

$$W = \frac{12S}{k^2(m^3 - m)} \tag{5}$$

The value of W ranges from 0 to 1, where $W = 1$ indicates complete agreement across all ranking rounds (implying the model's preference is highly stable), and $W = 0$ indicates no agreement. We utilize this metric to filter out instances where the model hallucinates or fails to distinguish quality differences effectively, thereby ensuring the robustness of the final averaged ranking.

To further guarantee the robustness of the constructed preference pairs and mitigate the potential biases inherent to a single evaluator, we implemented a rigorous multi-model verification pipeline. We employed an ensemble of three state-of-the-art multimodal large language models: Gemini-2.5-Pro, GPT-5, and GLM-4.5v.

Addressing the phenomenon of position bias—where models may exhibit a systematic preference for the first or

Table 1. Analysis of dataset quality under different filtering criteria. N denotes the number of samples. Corr., Err., and Controv. represent the percentages of samples classified as Correct, Error, and Controversial, respectively.

Method	N	Corr. (%)	Err. (%)	Controv. (%)
All Correct	20K	87.2	4.6	8.2
5 Correct + 1 Tie	33K	85.0	5.8	9.2
5 Correct + 1 Tie/Error	70K	78.6	10.7	10.7

second image in a sequence—we adopted a bidirectional evaluation strategy. For every candidate pair (I_A, I_B), each of the three models performed the evaluation twice: once with the original order and once with the image order swapped (i.e., (I_B, I_A)). This procedure yields a total of six independent assessments per sample. In each assessment, the models were tasked with a holistic evaluation covering fundamental integrity, typographical precision, and artistic quality, classifying the relationship as a clear preference, a tie, or a rejection of both images. This ensemble approach allows us to quantify the confidence level of each preference label based on the consensus among the six judgments.

To compare the reliability of our automated annotation framework under different filtering conditions, we manually annotate and analyze a subset of 1,000 samples from PosterPreference-70K. As described in Section 3, the final filtering stage involves three models, each performing two evaluations with the image order swapped, resulting in a total of six assessments per sample. We investigate three filtering strategies with varying levels of stringency: (1) retaining samples where all six model assessments are unanimous; (2) retaining samples with at least five consistent preference assessments, allowing for one tie; and (3) retaining samples with unanimous preference, allowing for one tie or one conflicting assessment. These samples are then annotated by four human experts and categorized into three groups: *Correct*, where the preference aligns with at least three annotators; *Error*, where the preference contradicts at least three annotators; and *Controversial* for all other cases.

As shown in Tab. 1, while the consistency between the automated filtering and human evaluations naturally decreases as the criteria become less stringent, the *5 Correct + 1 Tie/Error* strategy still maintains a robust alignment rate of 78.6%, with a relatively low error rate of 10.7%. This indicates that even under a slightly more relaxed filtering threshold, the data quality remains within an acceptable range for effective preference learning. Consequently, we elect to utilize the full 70K dataset for training. We posit that the significant increase in sample diversity and volume (from 20K/33K to 70K) outweighs the marginal gain in precision offered by stricter filtering, as larger-scale datasets typically drive better generalization capabilities in reward modeling. This 70K dataset is subsequently combined with a 100K preference pair subset from HPDv3 to form our final training corpus.

4. Ablation Studies on the Dataset Components

Table 2. Ablation study on dataset components. We compare the impact of using partial (33K) vs. full (70K) PosterPreference (PP) data, both independently and in combination with HPDv3. All values represent accuracy (↑).

Training Dataset Combination	HPDv3	PRB-Ad	PRB-Basic
PosterPreference-33K (Ours)	63.0	84.9	81.9
PosterPreference-70K (Ours)	63.9	84.6	83.0
HPDv3-100K	75.8	34.1	57.0
HPDv3 + PP-33K	76.9	84.5	80.8
HPDv3 + PP-70K (Final)	**77.1**	**85.0**	**83.9**

To validate our dataset selection strategy and quantify the contribution of different data sources, we conducted comprehensive ablation studies. Specifically, we investigate the performance of reward models trained on five distinct dataset configurations: (1) The high-consistency subset of our domain-specific dataset (**PP-33K**); (2) The full filtered domain-specific dataset (**PP-70K**); (3) The general aesthetic preference dataset (**HPDv3-100K**); (4) A combination of HPDv3 and the high-consistency subset (**HPDv3 + PP-33K**); and (5) A combination of HPDv3 and the full domain dataset (**HPDv3 + PP-70K**). We evaluated these models on the general HPDv3 test set as well as our domain-specific benchmarks, PRB-Ad and PRB-Basic.

As presented in Table 2, models trained solely on domain-specific data (PP-33K/70K) perform well on poster-related benchmarks but show limited generalization on the general HPDv3 test set. Conversely, training exclusively on HPDv3 yields strong general aesthetic judgment but suboptimal performance on poster-specific tasks (PRB-Ad/Basic), indicating a domain gap.

Crucially, combining the general and domain-specific datasets yields significant performance gains. Comparing the two combined settings, the model trained with the full **PP-70K** dataset consistently outperforms the one trained with the smaller **PP-33K** subset across all benchmarks. We also noted that the performance improvement was particularly significant on the out-of-domain test set PRB-Basic, indicating that more samples are beneficial for improving the generalization ability of the reward model. This empirical evidence supports our hypothesis in Section 3 that the benefits of increased data scale and diversity in the 70K dataset outweigh the marginal improvements in label consistency found in the 33K subset. Consequently, the *HPDv3 + PP-70K* configuration is adopted as our final training set, achieving the best balance between general aesthetic understanding and domain-specific expertise.

Figure 6. **Sample preference pairs from PosterRewardBench-Advanced.** This subset comprises images generated by Seedream and Qwen-Image-Lightning. It features complex poster layouts and supports both Chinese and English text rendering evaluations. The left side of each group shows the chosen image, the right side shows the rejected image, and the bottom displays an excerpt from the prompt.

5. Detailed Experimental Settings, Benchmark Construction

All experimental stages are conducted within the MS-Swift framework. We utilize DeepSpeed Zero-3 for Joint Supervised Fine-Tuning (Stage 1), Joint Rejection Sampling Fine-Tuning (Stage 2), and GRPO Fine-Tuning (Stage 4), while Zero-2 is employed for Scoring Module Training (Stage 3). Regarding hyperparameters, we set the per-device batch size to 1 with 2 gradient accumulation steps for the first two stages. For Stage 3, the batch size is set to 4 with 2 accumulation steps, whereas the GRPO stage uses a batch size of 1 with 8 accumulation steps. The adamw_torch optimizer is consistently applied across all stages.

PosterRewardBench. The PosterRewardBench comprises a total of 1,740 preference pairs. Specifically, PosterRewardBench-Advanced contains 1,223 pairs derived from Seedream 3.0, Seedream 4.0, and Qwen-Image-Lightning, while PosterRewardBench-Basic consists of 517 pairs sourced from Flux.1-dev, Flux.1-Krea-dev, and SD3.5-M. Within the Advanced subset, 488 pairs (39.9%) feature Chinese prompts and font rendering instructions, with the remainder comprising English counterparts. Conversely, given the incompatibility of Flux and SD3.5 with

Figure 7. **Sample preference pairs from PosterRewardBench-Basic.** This subset consists of images generated by Flux.1 and SD3.5 models. Due to model limitations, this benchmark focuses exclusively on English text rendering and standard graphic designs. The left side of each group shows the chosen image, the right side shows the rejected image, and the bottom displays an excerpt from the prompt.

the Chinese language, PosterRewardBench-Basic consists exclusively of English prompts. These pairs have been filtered from initial pools of 1,500 and 600 raw pairs, respectively. Four professional annotators evaluate preferences across multiple dimensions, and we retain only those pairs where at least three annotators reach a consensus. The data collection process spans approximately eight working days, with sample pairs illustrated in Figure 6 and Figure 7. In the pairwise evaluation, to verify decision stability regarding input order, we perform assessments with the "Chosen" image positioned both first and last. The specific prompts utilized are detailed in Figure 12. We also provide the prompt used by the analysis module when performing single-image analysis, as shown in Figure 11.

PosterBench. The PosterBench evaluation dataset consists of 250 prompts randomly sampled from PosterRewardBench, stratified into 100 cinematic and 150 non-cinematic themes. To rigorously assess compositional generalization, we employ distinct resolution strategies for each category. All cinematic prompts are standardized to a fixed vertical resolution of 832×1216 pixels (approximate aspect ratio of $2:3$), reflecting conventional movie poster standards. Conversely, the non-cinematic subset is designed to test adaptability to diverse geometries; each of the 150 prompts is assigned a unique target resolution (H_i, W_i) derived from its corresponding reference ground truth. Crucially, while target aspect ratios vary significantly *between* different prompts—spanning landscape, portrait, and square formats—the target resolution for any *specific* prompt remains constant across all evaluated models, ensuring a consistent baseline for comparative analysis.

Implementation of these target resolutions varies by model accessibility. Open-weights models (e.g., Qwen-Image[10], Flux[1]) are conditioned to generate images at the exact target dimensions (H_i, W_i) without resizing. However, for proprietary APIs with discrete resolution constraints, we employ an adaptive mapping strategy to approximate the target geometry. For Nano-Banana, arbitrary target dimensions are mapped to the nearest supported aspect ratio (e.g., $1:1, 2:3, 16:9$). Similarly, for GPT-Image-1, targets are bucketed into three presets: portrait (1024×1536) if the aspect ratio $\alpha = W/H < 0.8$, landscape (1536×1024) if $\alpha > 1.2$, and square (1024×1024) otherwise. This approach minimizes distortion while ensuring that closed-source models adhere

6. Limitation and Future Work

Despite the promising results achieved by our framework, several limitations remain to be addressed.

First, regarding the inherent position bias in MLLMs, we currently mitigate this issue in pairwise comparisons through a data balancing strategy. However, this strategy is not easily scalable to multi-image evaluation scenarios (e.g., ranking a batch of images simultaneously), where the combinatorial complexity of position swapping becomes prohibitive. Future work will focus on exploring more robust mechanisms to enhance the reliability and consistency of MLLMs in multi-image assessment tasks, reducing the dependency on permutation-based data augmentation.

Second, our proposed two-stage reward model, while offering superior interpretability and accuracy, entails a higher computational cost compared to standard discriminative reward models. The requirement for two inference steps inherently limits the model's real-time inference efficiency. In future research, we aim to optimize this architectural pipeline to reduce computational overhead. Additionally, we plan to conduct a comprehensive analysis of model scaling to investigate how different model sizes impact the trade-off between performance and efficiency in multimodal preference learning.

References

[1] Black Forest Labs. FLUX, 2024. https://github.com/black-forest-labs/flux.

[2] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[3] Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. Paddleocr 3.0 technical report, 2025.

[4] Maurice G Kendall and B Babington Smith. The problem of m rankings. *The annals of mathematical statistics*, 10(3): 275–287, 1939.

[5] Zongjie Li, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, and Yang Liu. Split and merge: Aligning position biases in llm-based evaluators. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11084–11108, 2024.

[6] Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15086–15095, 2025.

[7] Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in llm-as-a-judge. *arXiv preprint arXiv:2406.07791*, 2024.

[8] V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Jiazheng Xu, Jiale Zhu, Jiali Chen, Jing Chen, Jinhao Chen, Jinghao Lin, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Sheng Yang, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Shangqin Tu, Shengbiao Meng, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wenkai Li, Wei Jia, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyue Fan, Xuancheng Huang, Yanling Wang, Yadong Xue, Yanfeng Wang, Yanzi Wang, Yifan An, Yifan Du, Yiming Shi, Yiheng Huang, Yilin Niu, Yuan Wang, Yuanchang Yue, Yuchen Li, Yutao Zhang, Yuting Wang, Yu Wang, Yuxuan Zhang, Zhao Xue, Zhenyu Hou, Zhengxiao Du, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025.

[9] Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025.

[10] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025.

[11] 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. *arXiv preprint arXiv:2509.16117*, 2025.

Prompt 6.1 (MLLM Bo8 Prompt)

The following eight images were all generated by this Prompt:
{*prompt*}
Please compare these images and choose the one you think is the best to submit. Only the image number (1-8) will be returned.

Figure 8. **Prompt** for MLLM Ranking in Best-of-8 Selection.

Prompt 6.2 (Pointwise Analysis Collection Prompt)

You are a meticulous AI Image Quality Analyst and Creative Director. Your mission is to provide a detailed, structured analysis for a provided image based on the prompt.

Your Step-by-Step Analysis Process:

1. Deconstruct the prompt:

First, carefully read and internalize every detail of the provided prompt. Isolate its core components to establish your evaluation criteria.

2. Structured Image-by-Image Analysis:

For **the provided image**, you will conduct a methodical evaluation based on the following five criteria. Your analysis **MUST** address all five points in this specific order, analyzing both compliance and deviation within each point.

- **1. Fundamental Image Integrity:** Assess the image's foundational technical quality, independent of the prompt's creative instructions. Scrutinize for any objective flaws such as unintended blur, pixelation, overexposure, underexposure, or digital noise that detract from a professional finish.
- **2. AI Artifact and Realism Evaluation:** Examine the image specifically for common AI generation artifacts and assess its overall textural fidelity. If characters are present, conduct a rigorous check for anatomical inconsistencies, such as malformed hands, distorted facial features, or unnatural limb positioning. Evaluate whether the image achieves a believable, cohesive sense of realism or if artifacts disrupt the illusion.
- **3. Typographical Precision Analysis:** When the prompt includes text, perform a granular analysis of its execution. Evaluate adherence to the specified content, checking for spelling errors, omissions, or additions. Scrutinize the rendering of typographical details: font family, style, color, weight, and capitalization. Assess layout aspects like kerning, leading, alignment, and scale. Conclude by evaluating the overall legibility and seamless integration of the text.
- **4. Visual Prompt Interpretation:** Beyond text, assess the image's faithfulness to all other thematic and compositional directives in the prompt. Evaluate the accuracy of core elements (characters, objects, setting), the composition, layout, and adherence to the specified artistic style, mood, and color palette. Identify every element that aligns with the brief, as well as any deviations or creative misinterpretations.
- **5. Standalone Artistic Evaluation:** Disregarding the prompt's specific constraints, judge the image purely on its own artistic and aesthetic merits. Evaluate its composition, use of light and shadow, color theory, emotional impact, and overall visual appeal. Assess the technical execution and the creative choices made, determining if it is a compelling and well-crafted image in its own right.

The prompt is as follows:

"{creative_brief}"

Required Output:

Your final output must be structured as follows. Do not include any conversational introductions or summaries outside of this defined structure. Provide your structured analysis for the image. Use a clear heading for **Image Analysis**. Within this analysis, you must use the five numbered subheadings in the specified order. The total length should be within 500 words.

(Example of the final part of the output)

Image Analysis

1. Fundamental Image Integrity:

[Your analysis here.]

...

5. Standalone Artistic Evaluation:

[Your analysis here.]

Figure 9. **Prompt used for Pointwise Analysis Data Collection.**

Prompt 6.3 (Pairwise Preference Reasoning Prompt)

The images were generated by the prompt: "{prompt}".

Why is Image {better_image_index} better than Image {worse_image_index}?

When analyzing, you can consider these dimensions: Image Quality, AI Artifacts, Prompt Adherence, Text Rendering, and Aesthetic Value. Focus only on the key differentiating factors that make Image {better_image_index} superior. You don't need to cover all dimensions—only explain the core reasons.

Provide your analysis as a single paragraph.

Figure 10. **Prompt used for Pairwise Preference Reasoning Data Collection.**

Prompt 6.4 (Pointwise Analysis Generation Prompt)

Please analyze this image generated from the prompt: "{creative_brief}".
Provide a detailed analysis across these five dimensions:
1. Fundamental Image Integrity,
2. AI Artifact and Realism Evaluation,
3. Typographical Precision Analysis,
4. Visual Prompt Interpretation, and
5. Standalone Artistic Evaluation.

Figure 11. **Prompt used for Pointwise Analysis Generation Training and Inference.**

Prompt 6.5 (Pairwise Preference Prediction Prompt)

The following two images are generated from this prompt: "{prompt}".
Is Image 1 better than Image 2?
Please answer **Yes** or **No** first, then provide the reason.

Figure 12. **Prompt used for Pairwise Preference Prediction Training and Inference.**

Prompt 6.6 (Multi-Round Ranking Prompt (Part 1))

You are a meticulous AI Image Quality Analyst and Creative Director. Your mission is to provide a detailed, structured analysis for {image_ref} based on a creative brief, culminating in a definitive ranking presented in JSON format.

Your Step-by-Step Analysis Process:

1. Deconstruct the Creative Brief:
First, carefully read and internalize every detail of the provided prompt. Isolate its core components to establish your evaluation criteria.

2. Structured Image-by-Image Analysis:
For each of the {num_images} images, you will conduct a methodical evaluation based on the following five criteria. Your analysis for each image **MUST** address all five points in this specific order, analyzing both compliance and deviation within each point.

Crucially, your analysis for each image must be completely independent and self-contained. Do not make comparisons or references to any other image within an individual analysis block (e.g., in the analysis for Image 1, do not mention Image 2). All comparative logic is reserved for the final ranking synthesis.

- **1. Fundamental Image Integrity:** Assess the image's foundational technical quality, independent of the prompt's creative instructions. Scrutinize for any objective flaws such as unintended blur, pixelation, overexposure, underexposure, or digital noise that detract from a professional finish. Conversely, note the image's strengths in clarity, sharpness, and clean rendering.
- **2. AI Artifact and Realism Evaluation:** Examine the image specifically for common AI generation artifacts and assess its overall textural fidelity. If characters are present, conduct a rigorous check for anatomical inconsistencies, such as malformed hands, distorted facial features, or unnatural limb positioning. Evaluate whether the image achieves a believable, cohesive sense of realism or if artifacts disrupt the illusion.
- **3. Typographical Precision Analysis:** When the prompt includes text, perform a granular analysis of its execution. Evaluate adherence to the specified content, checking for spelling errors, omissions, or additions. Scrutinize the rendering of typographical details: font family, style, color, weight, and capitalization. Assess layout aspects like kerning, leading, alignment, and scale. Conclude by evaluating the overall legibility and seamless integration of the text.
- **4. Visual Prompt Interpretation:** Beyond text, assess the image's faithfulness to all other thematic and compositional directives in the prompt. Evaluate the accuracy of core elements (characters, objects, setting), the composition, layout, and adherence to the specified artistic style, mood, and color palette. Identify every element that aligns with the brief, as well as any deviations or creative misinterpretations.
- **5. Standalone Artistic Evaluation:** Disregarding the prompt's specific constraints, judge the image purely on its own artistic and aesthetic merits. Evaluate its composition, use of light and shadow, color theory, emotional impact, and overall visual appeal. Assess the technical execution and the creative choices made, determining if it is a compelling and well-crafted image in its own right.

Figure 13. **Prompt used for Multi-Round Ranking Data Collection (Part 1: Analysis Criteria).**

Prompt 6.7 (Multi-Round Ranking Prompt (Part 2))

3. Synthesize and Establish Ranking Logic:
After analyzing all images, synthesize your findings from the five-point analysis to establish a final ranking from best to worst. Your ranking **must** be a direct result of weighing your findings against this strict, **four-tier hierarchy of importance**:

- **Priority 1: Fundamental Image Integrity.** This is the primary gatekeeper for quality. An image must first be technically sound. Any image with significant fundamental flaws (e.g., pervasive blur, noise, or exposure issues) will be penalized heavily, regardless of its performance in other areas. This is evaluated in **point 1**.
- **Priority 2: Comprehensive Prompt Adherence.** Once an image passes the fundamental quality check, its faithfulness to the creative brief is the next most critical factor. This encompasses both textual accuracy and visual interpretation. Within this tier, accuracy in text is paramount. This is evaluated in **point 3 (Typographical Precision)** and **point 4 (Visual Prompt Interpretation)**.
- **Priority 3: AI-Generated Artifacts.** Images that are technically sound and prompt-adherent are then judged on their level of polish and realism. The absence of distracting AI-specific rendering errors, such as anatomical distortions or illogical object blending, is the third most important factor. This is evaluated in **point 2 (AI Artifact and Realism Evaluation)**.
- **Priority 4: Standalone Aesthetic Appeal.** This is the final consideration, used primarily to differentiate between images that perform similarly in the top three tiers. It is a subjective measure of the image's artistic merit, including composition, lighting, and overall impact. This is evaluated in **point 5**.

The Creative Brief (Prompt) is as follows:
"{creative_brief}"

Required Output:
Your final output must be structured as follows. Do not include any conversational introductions or summaries outside of this defined structure. First, provide your structured analysis for each of the {num_images} images. Use a clear heading for each image (e.g., **Image 1 Analysis**). Within each analysis, you must use the five numbered subheadings in the specified order. Following the analysis of all {num_images} images, provide the final ranking in a single, clean JSON block. The JSON block must be the very last thing in your response and contain only the ranking.

(Example of the final part of the output)
...
Image {num_images} Analysis
1. Fundamental Image Integrity:
[Your analysis here.]
...
5. Standalone Artistic Evaluation:
[Your analysis here.]

```json
{
"rank": {ranking_example}
}
```

Replace the numbers with the image numbers (1-{num_images}) in order from best to worst.

Figure 14. **Prompt used for Multi-Round Ranking Data Collection (Part 2: Synthesis and Output).**

Prompt 6.8 (MLLM Preference Assessment Prompt)

System Instruction: You are an AI Image Quality Analyst. Your task is to evaluate a pair of AI-generated images based on a creative brief and determine their data quality.

Evaluation Criteria:
Analyze each image based on its:
1. **Fundamental Image Integrity:** Technical quality (clarity, exposure, no major flaws).
2. **AI Artifacts and Realism:** Presence of AI artifacts, anatomical correctness, overall realism.
3. **Typographical Precision:** (If text is present) Accuracy of content, font, style, and layout.
4. **Visual Prompt Interpretation:** Faithfulness to the creative brief's theme, composition, style, and elements.
5. **Standalone Artistic Quality:** Aesthetic appeal, composition, lighting, and emotional impact.

Decision Task:
Based on a holistic evaluation of the criteria above, you must categorize the image pair into one of the following four options.

Output Options:
- **"Image 1"**: If Image 1 is clearly better than Image 2.
- **"Image 2"**: If Image 2 is clearly better than Image 1.
- **"Tie"**: If both images are good and of comparable quality.
- **"Both are bad"**: If both images significantly fail the prompt's core requirements. This includes, but is not limited to: major errors in text or subject matter, missing key content, severe image quality issues, or extremely low aesthetic value.

The Creative Brief (Prompt):

"{creative_brief}"

Required Output Format:
Your final output must contain **only one** of the four decision strings and nothing else. Do not provide analysis, justifications, or any other conversational text.

Example:
Image 1

Figure 15. **Prompt used for MLLM Pairwise Preference Verification.** This prompt guides the model to evaluate image pairs across five dimensions, specifically filtering out low-quality samples via the "Both are bad" option.