WHEN PREFERENCE MARGINS HIDE TEXT-ONLY SHORTCUTS: VISUAL-CONDITIONAL MARGIN LEARNING FOR MULTIMODAL REWARD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal reward models (MM-RMs) should assess response quality using cross-modal information, yet often rely on text-only shortcuts, leading to underuse of visual evidence and potentially impairing reward judgments and downstream optimization. By analyzing text-only preference signals in pretrained representations and their evolution during reward-model training, we find that the standard Bradley–Terry (BT) objective widely used to train reward models from pairwise preferences, despite improving multimodal preference prediction, further amplifies pre-existing textual preference signals and leaves the contribution of visual evidence to the pairwise preference margin unsupervised. Based on these observations, we propose visual-conditional margin learning (VCM), a unified framework with two levels of visual-conditional supervision: visual-gain supervision from text-only to multimodal conditioning, and stronger evidence-flip (EF) supervision enforcing preference-margin changes under local reversals of decisive visual evidence while keeping the question and responses fixed, thereby cancelling the shared text-only component. We evaluate our method across multiple MM-RM backbones and multiple multimodal preference benchmarks, showing improved reward-model performance, reduced text-only shortcut dependence, and better downstream response selection in Best-of-N evaluation.

1 INTRODUCTION

Reward models (RMs) provide preference signals for aligning language models with human preferences (Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022). Multimodal reward models (MM-RMs) incorporate multimodal context into preference modeling, as response quality may depend on cross-modal evidence unavailable from text alone (Sun et al., 2024; Zang et al., 2025; Zhang et al., 2026; Jin et al., 2026). Reliable MM-RMs are important for applications including visual hallucination detection, multimodal safety evaluation, and multimodal reasoning assessment (Li et al., 2025a; Yasunaga et al., 2025).

Despite recent progress, MM-RMs remain susceptible to unimodal spurious correlations in multimodal preference data (Li et al., 2025b). A prominent example is the *text-only shortcut*: textual patterns in the question–response context can correlate with preference labels, allowing an MM-RM to make correct predictions without sufficiently integrating cross-modal evidence. As illustrated in Figure 1, high pairwise preference accuracy can mask substantial reliance on textual shortcuts, such that strong performance alone does not establish that the learned preference is genuinely grounded in multimodal evidence. When these shortcut correlations no longer hold, the resulting preference judgments can become unreliable. Because reward models provide preference signals for downstream alignment and response selection, such shortcut dependence can further affect downstream optimization.

To understand how this dependence arises, we examine text-only preference signals before and during reward-model training. We find that pretrained VLM representations already contain preference-related signals that remain predictive even without visual inputs. Moreover, although the standard Bradley–Terry (BT) objective used to train reward models from pairwise preferences improves multimodal preference accuracy, it further strengthens the reliance on these pre-existing text-only sig-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

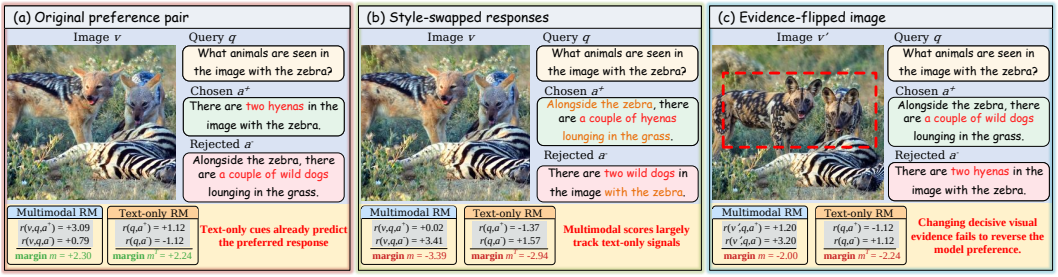


Figure 1: **Illustrative failure modes of text-only dependence in MM-RMs. (a)** Text-only scoring already predicts the original preference. **(b)** Altering superficial textual cues, such as response patterns or length, reverses both multimodal and text-only margins. **(c)** Editing decisive visual evidence fails to reverse the model preference. Here, the margin is the reward-score difference between the preferred and rejected responses; m and m^T denote multimodal and text-only preference margins.

nals. This reveals a limitation of standard pairwise supervision: BT only constrains the overall preference margin, i.e., the reward-score difference between two responses, without specifying how much of this difference is attributable to visual conditioning. Therefore, accurate multimodal preference prediction alone does not establish that the learned preference is grounded in visual evidence, motivating explicit supervision of the visual contribution to the preference margin.

Based on these observations, we propose visual-conditional margin learning (VCM), a unified framework with two levels of visual-conditional supervision. Visual-gain supervision compares multimodal and text-only states and requires visual conditioning to provide an additional preference margin in the ground-truth direction. Stronger evidence-flip (EF) supervision locally reverses the decisive visual evidence while keeping the question and responses fixed, requiring the preference margin to change accordingly. The resulting cross-state contrast cancels the shared text-only component, providing a direct optimization signal for preference changes induced by visual evidence. Both supervision terms are jointly optimized with standard BT in a single training stage.

The main contributions of this paper are:

- We analyze text-only preference signals in MM-RMs before and during reward-model training, showing that pretrained VLM representations already encode preference-related textual signals and that standard BT training further strengthens reliance on these pre-existing signals.
- We propose visual-conditional margin learning (VCM), a unified framework that strengthens visual contributions to the preference margin at two levels: visual-gain supervision from text-only to multimodal conditioning, and stronger evidence-flip (EF) supervision under controlled reversals of decisive visual evidence.
- We conduct extensive evaluations across multiple MM-RM backbones and multimodal preference benchmarks, showing consistent performance gains, reduced reliance on text-only shortcuts and unimodal spurious correlations, and better Best-of-N response selection.

2 Related Work

Multimodal Reward Modeling. Multimodal reward modeling has become an important component of multimodal alignment and evaluation, enabled by preference datasets such as RLHF-V (Yu et al., 2024a), VLFeedback (Li et al., 2024) and RLAIF-V (Yu et al., 2025). Building on these data, recent work has developed scalar MM-RMs for multimodal response evaluation (Wang et al., 2025a; Zang et al., 2025; Wang et al., 2025b; Zhang et al., 2026). Beyond assigning a single scalar score to each response, TLDR (Fu et al., 2025a) provides token-level reward annotations for fine-grained response assessment, while VL-MDR (Chen et al., 2026) decomposes evaluation into input-dependent dimensions. In parallel, benchmarks such as VL-RewardBench (Li et al., 2025a) provide targeted evaluation of multimodal reward models across visual perception, hallucination, and reasoning. Existing work has primarily advanced reward-model design, richer supervision, and broader multimodal evaluation (Yan et al., 2025; Jin et al., 2026). Our work instead studies the contribution of multimodal conditioning to the learned pairwise preference margin beyond text-only signals.

Unimodal Shortcuts and Multimodal Preference Learning. Preference and reward models can exploit superficial features correlated with preference labels, motivating methods for mitigating spurious preference cues (Shen et al., 2023; Chen et al., 2024b; Huang et al., 2025; Bharadwaj et al., 2026). In multimodal reward modeling, Li et al. (2025b) show that MM-RMs can exploit unimodal spurious correlations, particularly text-only shortcuts, and propose shortcut-aware learning that reweights examples using a text-only proxy. Related work on generative MLLM judges identifies perceptual judgment bias and improves visual grounding through training on perceptually perturbed responses (Park et al., 2026). At the policy level, mDPO (Wang et al., 2024) and CHiP (Fu et al., 2025b) introduce visual preference objectives for multimodal preference optimization, while S-VCO (Wu et al., 2025) and JoMPO (Kim & Yoon, 2026) leverage contrastive visual contexts to strengthen fine-grained visual alignment. Recent PEA-DPO (Feng et al., 2026) further targets visual insensitivity with explicit visual preference signals. Complementary to these efforts, our work formulates the issue in terms of modality-specific contribution to the learned pairwise margin, focusing on preference-margin changes induced by multimodal conditioning beyond text-only signals.

3 Preliminary Study

To analyze the presence and evolution of text-only dependence in MM-RMs, we conduct a systematic study along two dimensions: (1) the presence of text-only preference signals in pretrained VLM representations; and (2) the dynamics of these signals under standard Bradley–Terry (BT) training.

Throughout this paper, we denote a multimodal preference example by (x, y), where $x = (I, Q, a_1, a_2)$ contains an image I, a question Q, and two responses a_1 and a_2. The label $y \in \{+1, -1\}$ specifies the preferred response, with $y = +1$ for $a_1 \succ a_2$ and $y = -1$ for $a_2 \succ a_1$. An MM-RM $r_\theta(I, Q, a)$ assigns a scalar reward to a. We define the preference margin as

$$m_\theta = r_\theta(I, Q, a_1) - r_\theta(I, Q, a_2). \tag{1}$$

Standard Bradley–Terry (BT) training (Bradley & Terry, 1952) minimizes

$$\mathcal{L}_{\mathrm{BT}} = -\mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\log \sigma(y m_\theta)\right]. \tag{2}$$

We report pairwise accuracy for preference prediction.

We conduct the analysis with Qwen2.5-VL-7B-Instruct (Bai et al., 2025b) and InternVL3-8B (Zhu et al., 2025) on POVID (Zhou et al., 2024), RLAIF-V (Yu et al., 2025), and RLHF-V (Yu et al., 2024a). POVID constructs preference pairs via hallucination injection, RLAIF-V uses AI-generated feedback across diverse tasks and domains, and RLHF-V derives preferences from fine-grained human corrections. From each dataset, we sample 3,000 training and 1,000 item-disjoint test pairs.

3.1 Text-Only Preference Signals in Pretrained Representations

We first examine whether pretrained VLM representations already contain preference-related signals before any reward-specific optimization. We consider two input settings: in the multimodal setting, each response is encoded with the original image and question; in the text-only setting, we remove the image while keeping the question and response unchanged. For each response a_i, we extract the hidden state of its final response token, denoted by h_i^{M} and h_i^{T} for the two settings.

We train only a linear probe (Belinkov, 2022) on the representation difference between the two responses. For each setting $c \in \{\mathrm{M}, \mathrm{T}\}$, the probe defines the pairwise margin

$$m_{\mathrm{probe}}^c = w_c^\top (h_1^c - h_2^c), \tag{3}$$

optimized with the same pairwise BT objective as the reward model. We fit the probes on the training split and evaluate them on the item-disjoint test split.

As shown in Figure 2(a), text-only probes achieve above-chance preference accuracy in most settings, while the absolute accuracy varies markedly across datasets. More importantly, text-only performance closely tracks the corresponding multimodal probes, with only a modest gap across datasets and backbones. These results show that pretrained VLM representations already encode preference-related textual signals before reward modeling. The small gap between text-only and multimodal probes indicates that much of the apparent multimodal preference predictability is already available from text alone, providing a strong shortcut that subsequent reward modeling can exploit without sufficiently relying on visual evidence.

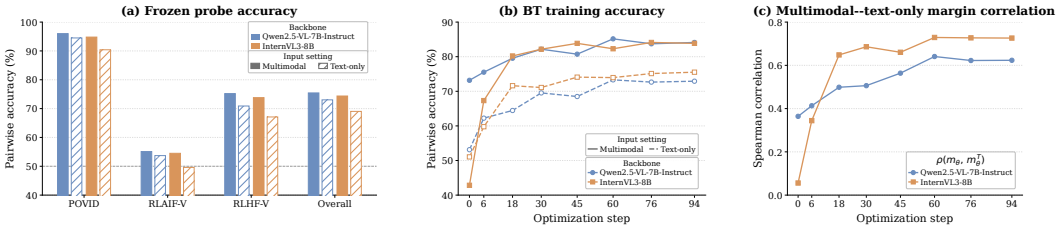


Figure 2: **Preliminary analysis of text-only dependence in MM-RMs.** We evaluate Qwen2.5-VL-7B-Instruct and InternVL3-8B under multimodal and text-only settings. **(a)** Frozen probes reveal substantial text-only preference predictability in pretrained representations. **(b)** Preference accuracy during standard BT training, where text-only accuracy increases alongside multimodal accuracy. **(c)** Spearman correlation between multimodal and text-only preference margins, showing increasing alignment between multimodal and text-only preference margins during training.

These dataset-level differences are consistent with variations in preference-pair construction, such as response length, style, and hallucination-oriented patterns. Together, these results identify a pre-existing source of text-only dependence: pretrained VLM representations already encode textual features correlated with preference labels before reward modeling, creating the possibility that subsequent BT optimization further amplifies this dependence.

3.2 Text-Only Preference Dynamics during BT Training

We next examine how these pre-existing text-only preference signals evolve during reward modeling. Starting from the same pretrained VLM backbones, we replace the language-modeling head with a linear reward head and train MM-RMs on multimodal preference pairs using standard BT.

For training-dynamics analysis, we evaluate selected optimization steps on 256 held-out pairs from each dataset under both multimodal and text-only settings. For a trained MM-RM, we define the text-only preference margin as

$$m_\theta^T = r_\theta(\varnothing, Q, a_1) - r_\theta(\varnothing, Q, a_2). \tag{4}$$

We track pairwise accuracy under both input settings, together with the Spearman correlation $\rho(m_\theta, m_\theta^T)$ between multimodal and text-only preference margins.

As shown in Figure 2(b), standard BT training steadily improves multimodal preference accuracy as expected. Meanwhile, text-only accuracy increases alongside multimodal accuracy, even without explicit text-only supervision. This shows that BT optimization further amplifies the pre-existing textual preference signal, making it increasingly influential in preference prediction.

Figure 2(c) shows a concurrent increase in $\rho(m_\theta, m_\theta^T)$, indicating that the relative ordering of preference margins under multimodal and text-only inputs becomes increasingly similar during training. This growing alignment suggests that multimodal preference margins increasingly reflect preference structure already present in the text-only margins, making visual contributions harder to distinguish and increasing the risk of underusing cross-modal information.

Collectively, these analyses show that text-only preference signals are already present in pretrained VLM representations and are further amplified during standard BT training. This exposes a limitation of standard pairwise supervision: BT constrains only the overall preference margin, without specifying how much of that margin is attributable to visual conditioning.

4 Visual-Conditional Margin Learning

Accordingly, we introduce *visual-conditional margin learning* (VCM), a unified framework that explicitly supervises the visual contribution to the preference margin at two levels. *Visual-gain supervision* encourages multimodal conditioning to increase the margin over the text-only state in the ground-truth direction, while stronger *evidence-flip supervision* constrains margin changes under controlled reversals of decisive visual evidence. Both are jointly optimized with standard BT in a single training stage, as illustrated in Figure 3.

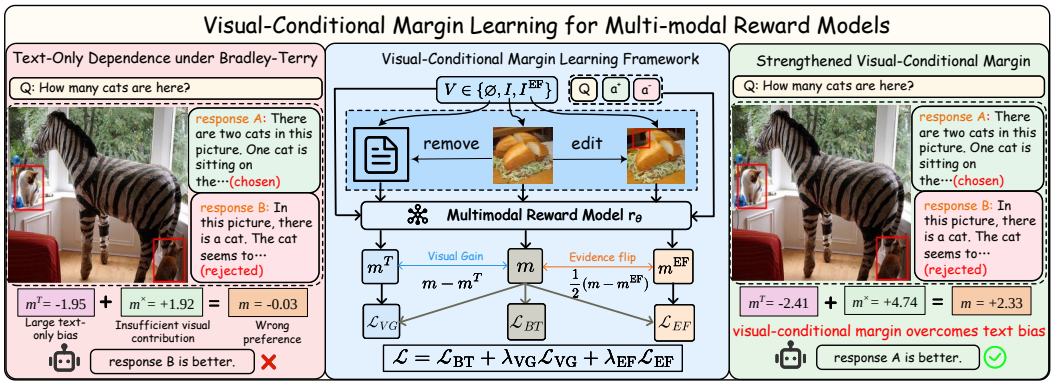


Figure 3: **Overview of visual-conditional margin learning (VCM).** VCM supervises preference-margin changes from text-only to multimodal conditioning and under evidence-flip contrasts. The preference margin is the reward-score difference between two responses. Here, m, m^T, and m^{EF} denote the multimodal, text-only, and evidence-flip margins, respectively, while $m^\times = m - m^T$ denotes the visual-conditional contribution.

4.1 Visual-Conditional Margin Decomposition

Standard BT supervises only the overall preference margin m_θ, without constraining whether this margin is supported by cross-modal interaction or text-only cues. When text-only signals already favor the preferred response, BT can further enlarge the preference margin through the text-only pathway without strengthening the desired cross-modal interaction.

To characterize the contribution beyond text-only scoring, we use the text-only reward as a reference and define $r_\theta^\times(I, Q, a) = r_\theta(I, Q, a) - r_\theta(\varnothing, Q, a)$. The preference margin can then be written as $m_\theta = m_\theta^T + m_\theta^\times$, where

$$m_\theta^\times = r_\theta^\times(I, Q, a_1) - r_\theta^\times(I, Q, a_2) = m_\theta - m_\theta^T. \tag{5}$$

Here, $m_\theta^\times$ captures the incremental pairwise preference margin associated with cross-modal interaction beyond text-only scoring. Accordingly, $m_\theta - m_\theta^T$ provides a direct proxy for supervising the cross-modal contribution to preference prediction.

This motivates supervising preference-margin differences across conditioning states, from text-only to multimodal inputs and under stronger evidence-flip contrasts.

4.2 Visual-Gain Supervision

We use $m_\theta - m_\theta^T$ to quantify the preference-margin change introduced by multimodal conditioning, which we refer to as the *visual gain*. Multiplying by the preference label y makes a positive value correspond to a change in the ground-truth preference direction. We require this gain to exceed a threshold $\gamma_{\mathrm{VG}} > 0$ using a squared hinge objective.

A direct objective on $m_\theta - m_\theta^T$ also backpropagates through m_θ^T, which can satisfy the constraint by pushing the text-only margin against the preference direction. This is undesirable because our goal is to strengthen the cross-modal contribution rather than explicitly suppress text-only preference signals. We therefore retain only a fraction β of the gradient through m_θ^T:

$$\mathcal{L}_{\mathrm{VG}} = \mathbb{E}\left[\gamma_{\mathrm{VG}} - y\left(m_\theta - \left[\beta m_\theta^T + (1-\beta)\,\mathrm{sg}(m_\theta^T)\right]\right)\right]_+^2, \tag{6}$$

where $[z]_+ = \max(z, 0)$ and $\mathrm{sg}(\cdot)$ denotes stop-gradient, which preserves the forward value but blocks gradient backpropagation. Here, γ_{VG} specifies the visual-gain threshold, while $\beta \in [0, 1]$ controls the gradient passed through m_θ^T. Thus, the forward value remains $m_\theta - m_\theta^T$, while a smaller β reduces reverse optimization of the text-only margin. Once the gain reaches γ_{VG}, the hinge term becomes inactive.

4.3 Evidence-Flip Supervision

As a stronger form of visual-conditional supervision, we construct *evidence-flip* (EF) pairs that provide a controlled contrast over the visual evidence determining the preference. Given an original example (I, Q, a_1, a_2, y), we locally edit the decisive visual evidence to obtain I^{EF} while keeping Q, a_1, a_2 unchanged, with the edit intended to reverse the preference from y to $-y$. Details of EF construction are provided in Appendix B.

Let the preference margin under the edited image be $m_\theta^{\mathrm{EF}} = r_\theta(I^{\mathrm{EF}}, Q, a_1) - r_\theta(I^{\mathrm{EF}}, Q, a_2)$. Because the original and edited examples share the same question and responses, their text-only margins are identical, while their preference labels are y and $-y$, respectively. Using the decomposition from Section 4.1,

$$\frac{1}{2} y \left(m_\theta - m_\theta^{\mathrm{EF}}\right) = \frac{1}{2} \left[y m_\theta^{\times} + (-y) m_\theta^{\times,\mathrm{EF}} \right]. \tag{7}$$

Thus, the cross-state difference exactly cancels the shared text-only component and directly supervises the cross-modal preference contributions under the two opposing visual states. A purely text-only MM-RM cannot satisfy this constraint, as it assigns identical margins to the two states with opposite labels and achieves chance-level accuracy over the paired original–EF examples. EF therefore places direct training pressure on the preference margin to respond to the locally altered visual evidence.

We supervise this cross-state margin change with

$$\mathcal{L}_{\mathrm{EF}} = \mathbb{E}\left[\gamma_{\mathrm{EF}} - \frac{1}{2} y \left(m_\theta - m_\theta^{\mathrm{EF}}\right) \right]_+^2, \tag{8}$$

where $\gamma_{\mathrm{EF}} > 0$ specifies the required margin change between the original and evidence-flipped states. The factor $1/2$ normalizes the contribution from the two opposing visual states. Since the shared text-only component is canceled in the cross-state difference, minimizing $\mathcal{L}_{\mathrm{EF}}$ requires the preference margin to change according to the altered visual evidence rather than relying on image-invariant textual shortcuts.

4.4 Overall Objective

We jointly optimize standard BT on the original preference pairs with visual-conditional supervision, including visual-gain supervision on perception-dependent examples and evidence-flip supervision on their corresponding edited pairs. The overall objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{BT}} + \lambda_{\mathrm{VG}} \mathcal{L}_{\mathrm{VG}} + \lambda_{\mathrm{EF}} \mathcal{L}_{\mathrm{EF}}, \tag{9}$$

where λ_{VG} and λ_{EF} control the strengths of the two visual-conditional supervision terms.

All objectives are optimized jointly in a single training stage, combining standard preference optimization with explicit supervision of visual-conditional preference margins.

5 Experiments

We evaluate our VCM learning framework on multimodal preference benchmarks, further analyzing shortcut behavior, visual-conditional margins, and downstream Best-of-N response selection.

5.1 Experimental Setup

Training data and models. We construct a 6K preference set from POVID (Zhou et al., 2024), RLAIF-V (Yu et al., 2025), and RLHF-V (Yu et al., 2024a) (approximately 3K/2K/1K), with a larger POVID portion to increase exposure to the stronger unimodal spurious correlations observed in our preliminary study. Standard BT is applied to all preference pairs; since preferences in POVID and RLHF-V are primarily determined by visual evidence, we partition the corresponding training pairs into two disjoint visual-conditional supervision subsets: one receives visual-gain supervision, while the other receives the stronger evidence-flip supervision with additionally constructed edited images. We evaluate Qwen2.5-VL-7B-Instruct (Bai et al., 2025b), Qwen3-VL-8B (Bai et al., 2025a), and InternVL3-8B (Zhu et al., 2025) as reward-model backbones, freezing the vision encoder and

Table 1: Main reward-modeling results across three vision-language backbones. BT and SA are baseline objectives, while VCM denotes our full method. −EF and −VG are ablations of VCM. Best and second-best results within each backbone are shown in **bold** and underlined, respectively.

Backbone	Method	Held-out↑	VLF↑	WildVis↑	VLRB-H↑	MMRB↑	MLLM-Pair↑
Qwen2.5-VL-7B	BT	80.91±0.30	65.57±0.61	63.10±1.30	87.32	68.73	56.48
	SA	79.69±0.88	73.29±1.02	64.37±2.05	86.25	73.67	60.10
	VCM	**81.83**±0.71	**81.88**±0.81	**70.50**±1.25	86.52	**76.03**	**64.49**
	−EF	80.46±0.56	76.53±0.54	65.47±1.39	**89.19**	72.90	63.07
	−VG	79.12±0.60	79.75±0.66	69.10±1.08	85.05	75.86	61.46
Qwen3-VL-8B	BT	82.00±0.44	73.76±0.16	66.13±1.07	**89.19**	74.66	56.84
	SA	**82.16**±0.73	71.78±0.41	64.93±1.61	85.45	**78.22**	60.63
	VCM	81.49±0.86	**80.56**±0.73	**70.10**±1.35	88.25	78.11	**63.63**
	−EF	81.82±0.74	79.06±0.79	67.90±1.40	87.72	74.44	59.74
	−VG	81.44±0.98	79.56±0.56	**70.10**±2.30	87.05	75.75	61.98
InternVL3-8B	BT	**82.44**±0.93	71.80±0.56	67.13±0.67	87.32	63.58	54.24
	SA	80.90±0.95	65.21±0.91	63.60±1.40	85.18	65.44	58.43
	VCM	81.10±0.99	**74.15**±1.20	**67.93**±0.95	85.45	**65.83**	**60.99**
	−EF	82.32±1.05	74.13±0.52	66.67±0.90	**87.72**	62.92	58.80
	−VG	78.67±0.92	73.86±0.11	67.10±0.66	82.78	55.18	60.73

connector while optimizing language LoRA parameters (Hu et al., 2022) and a scalar reward head. Additional training and implementation details are provided in Appendix A.

Baselines and protocol. We compare standard BT, Shortcut-aware (SA) learning (Li et al., 2025b), and our full VCM framework. SA uses a text-only auxiliary branch to compute the Shortcut-Failure Coefficient (SFC), downweighting examples that are well explained without the image. We further include two ablations: VCM-EF removes the edited EF images and $\mathcal{L}_{\mathrm{EF}}$, replacing EF supervision with VG on the corresponding original pairs, while VCM-VG removes $\mathcal{L}_{\mathrm{VG}}$ and retains EF supervision. All methods otherwise follow the same backbone and optimization protocol. Additional implementation and ablation details are provided in Appendix A.

Evaluation. We evaluate on multimodal RM benchmarks covering visual perception, hallucination, and general multimodal correctness. For POVID, RLAIF-V, and RLHF-V, we construct 1K held-out test sets with strict train–test deduplication and report their aggregate performance as *Held-out*. We use three 1K-instruction VLFeedback rolls (Li et al., 2024), expanded into non-tied pairs, and three 1K-pair WildVision-Battle rolls (Lu et al., 2024). Results on Held-out, VLFeedback, and WildVision-Battle are averaged over three rolls. We further evaluate on fixed public benchmarks, including VL-RewardBench (VLRB) (Li et al., 2025a), Multimodal RewardBench (MMRB) (Yasunaga et al., 2025), and MLLM-as-a-Judge Pair (Chen et al., 2024a). For VLRB, we focus on its hallucination subset, which shares source datasets with our training data but is strictly deduplicated. For MMRB, we report the aggregate over VQA and general correctness. Pairwise accuracy is the primary metric. Inference uses only the original multimodal branch and requires no auxiliary branch.

5.2 Main Results

As shown in Table 1, the full VCM framework delivers strong and consistent gains across all three vision-language backbones, with the clearest improvements on external preference benchmarks. VCM achieves the best VLFeedback and MLLM-Pair results for all three backbones, and the best or tied-best WildVision performance. Compared with SA, VCM improves VLFeedback by 8.59–8.94 points across backbones; compared with standard BT, it improves MLLM-Pair by 6.75–8.01 points. VCM also consistently improves MMRB over BT while remaining competitive on Held-out and VLRB-H. Overall, these results show that visual-conditional margin supervision generalizes across

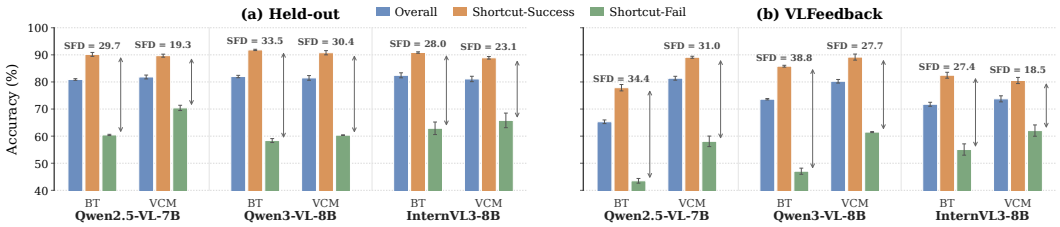


Figure 4: Shortcut-behavior analysis on **(a)** Held-out and **(b)** VLFeedback across three backbones. We report overall multimodal accuracy and accuracy on shortcut-success and shortcut-fail subsets. SFD measures the accuracy gap between the two subsets, with lower values indicating weaker dependence on text-only shortcuts.

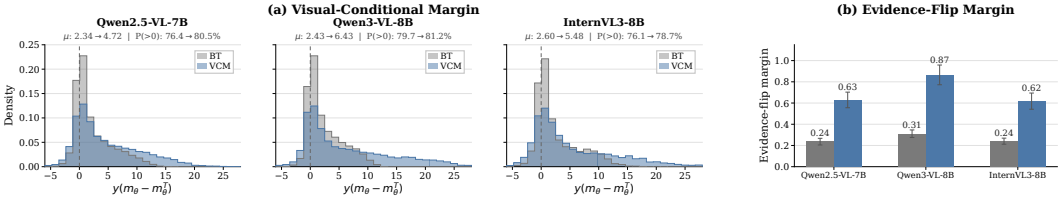


Figure 5: Margin analysis across three backbones under BT and VCM. **(a)** Distributions of visual-conditional margins. **(b)** Evidence-flip margins.

backbone architectures and improves preference generalization beyond standard BT and shortcut-aware reweighting.

The ablations further show complementary roles of the two supervision terms. Removing evidence-flip supervision ($-$EF) consistently weakens performance on VLFeedback, WildVision, MMRB, and MLLM-Pair across most backbones, consistent with the stronger visual contrast provided by controlled evidence reversals. Removing visual-gain supervision ($-$VG) causes broader degradation relative to the full model, including on Held-out and VLRB-H. These results suggest that visual-gain supervision provides a general constraint on cross-modal preference contribution, while evidence-flip supervision further strengthens sensitivity to decisive visual evidence; combining both yields the most robust overall performance.

5.3 Analysis of Shortcut Behavior

We further examine whether VCM reduces reliance on text-only shortcuts. Following Li et al. (2025b), we partition examples according to text-only preference correctness: the *shortcut-success* subset contains examples correctly predicted without the image, while the *shortcut-fail* subset contains those where the text-only prediction fails. Let T and M denote correctness under text-only and multimodal inputs, respectively. Shortcut-Failure Degradation (SFD) is defined as $\mathrm{SFD} = \Pr(M = 1 \mid T = 1) - \Pr(M = 1 \mid T = 0)$. A large SFD indicates that multimodal preference accuracy depends strongly on whether the text-only shortcut succeeds.

As shown in Figure 4, VCM consistently reduces SFD relative to standard BT across all three backbones on both Held-out and VLFeedback. On Held-out, SFD decreases by 3.1–10.4 points across backbones; on VLFeedback, the reduction reaches 3.4–11.1 points. Importantly, VCM consistently improves accuracy on the shortcut-fail subset, increasing $\Pr(M = 1 \mid T = 0)$ across backbones. This means that when text-only cues lead to an incorrect preference, visual conditioning is more likely to correct the prediction, directly indicating weaker reliance on text-only shortcuts and more effective use of visual evidence.

5.4 Visual-Conditional Margin Analysis

We further examine whether VCM directly strengthens the preference contribution induced by visual conditioning. We analyze the visual-conditional margin $y(m_\theta - m_\theta^T)$, where a positive value in-

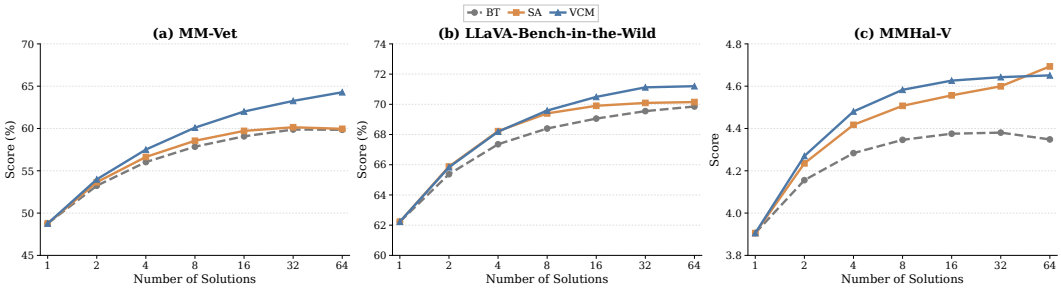


Figure 6: Best-of-N selection across benchmarks, averaged over three reward-model backbones.

dicates that adding the image increases the preference margin in the ground-truth direction. We also measure the evidence-flip margin $\frac{1}{2}y(m_\theta - m_\theta^{\mathrm{EF}})$, which quantifies the preference-margin change under a controlled reversal of decisive visual evidence.

As shown in Figure 5(a), VCM consistently shifts the visual-conditional margin toward larger positive values across all three backbones. The mean margin increases from 2.34–2.60 under BT to 4.72–6.43 under VCM, while the fraction of positive margins increases from 76.1–79.7% to 78.7–81.2%. This shows that multimodal conditioning provides a stronger additional preference signal in the correct direction under VCM. Figure 5(b) shows the same trend under stronger evidence-flip supervision: the average margin increases from 0.24–0.31 under BT to 0.62–0.87 under VCM. Together, these results confirm that VCM strengthens both the incremental contribution of cross-modal information and the model's preference response to controlled changes in decisive visual evidence.

5.5 Best-of-N Selection

We evaluate the downstream utility of the learned RMs through Best-of-N (BoN) selection (Gao et al., 2023) on MM-Vet (Yu et al., 2024b), LLaVA-Bench-in-the-Wild (Liu et al., 2023), and MMHal-V (Sun et al., 2024; Amirloo et al., 2024), covering general multimodal capabilities, open-ended instruction following and reasoning, and hallucination-sensitive response quality. For each of the 374 image–query pairs, a frozen InternVL2-8B (OpenGVLab, 2024) generates a shared pool of 64 candidate responses. Each reward model scores the same candidates and selects the highest-scoring response from $N \in \{1, 2, 4, 8, 16, 32, 64\}$. For each N, we report the exact expectation over uniformly sampled candidate subsets and evaluate the selected responses with a fixed GPT-5.4 mini multimodal judge (OpenAI, 2026a). We compare reward models trained with BT, SA, and VCM.

As shown in Fig. 6, all methods coincide at $N = 1$, while their differences become increasingly apparent as the candidate pool grows. VCM achieves consistently stronger selection performance on MM-Vet and LLaVA-Bench-in-the-Wild, particularly at larger N, while remaining competitive on MMHal-V. These results show that the improved visual preference signal learned by VCM translates into more effective downstream response selection as more candidate responses become available.

6 Conclusion

This paper studies text-only dependence in multimodal reward models and shows that pretrained VLM representations already contain text-only preference signals that are further amplified during standard BT training. These findings reveal a limitation of standard pairwise supervision: improving the overall preference margin does not ensure sufficient contribution from cross-modal information. To address this issue, we introduce visual-conditional margin learning (VCM), which explicitly supervises the visual-conditional contribution to the preference margin through visual-gain and stronger evidence-flip supervision. Experiments across multiple reward-model backbones and multimodal preference benchmarks demonstrate consistent improvements over standard BT and shortcut-aware learning. Further analyses show reduced text-only shortcut dependence, stronger visual-conditional and evidence-flip margins, and improved robustness when textual shortcuts fail. Best-of-N evaluation further shows that these improvements translate into more effective downstream response selection.

Reproducibility Statement

We provide detailed reproducibility information in the appendices. Appendix A describes the training data, model and optimization settings, baseline configurations, evaluation protocols, and Best-of-N setup, while Appendix B details the construction and inspection of evidence-flip samples. The provided implementation and documented protocols are intended to support reproduction of the reported experiments and experimental trends.

AI Use Statement

We used generative AI tools in the construction, evaluation, and preparation of this work. GPT Image 2 was used to generate local image edits for evidence-flip samples, with the construction and inspection procedure described in Appendix B. GPT-5.4 mini was used as the fixed multimodal evaluator for the Best-of-N experiments in Section 5.5 and Appendix A.5. Generative AI tools were also used to assist with manuscript organization and language editing. All AI-assisted data, evaluations and manuscript content included in the final paper were reviewed by the authors.

References

Elmira Amirloo, Jean-Philippe Fauconnier, Christoph Roesmann, Christian Kerl, Rinu Boney, Yusu Qian, Zirui Wang, Afshin Dehghan, Yinfei Yang, Zhe Gan, and Peter Grasch. Understanding Alignment in Multimodal LLMs: A Comprehensive Study, 2024. URL https://arxiv.org/abs/2407.02477.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL Technical Report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report, 2025b. URL https://arxiv.org/abs/2502.13923.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, 2022. URL https://arxiv.org/abs/2204.05862.

Yonatan Belinkov. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1):207–219, March 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Anirudh Bharadwaj, Chaitanya Malaviya, Nitish Joshi, and Mark Yatskar. Flattery, Fluff, and Fog: Diagnosing and Mitigating Idiosyncratic Biases in Preference Models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 19171–19191, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/2096bafd3073a2224f6f0adb594068df-Paper-Conference.pdf.

Ralph Allan Bradley and Milton E. Terry. RANK ANALYSIS OF INCOMPLETE BLOCK DESIGNS: THE METHOD OF PAIRED COMPARISONS. *Biometrika*, 39(3-4):324–345, 12 1952. ISSN 0006-3444. doi: 10.1093/biomet/39.3-4.324. URL https://doi.org/10.1093/biomet/39.3-4.324.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. MLLM-as-a-Judge: Assessing Multimodal LLM-as-a-Judge with Vision-Language Benchmark. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6562–6595. PMLR, 21–27 Jul 2024a. URL https://proceedings.mlr.press/v235/chen24h.html.

Lichang Chen, Chen Zhu, Jiuhai Chen, Davit Soselia, Tianyi Zhou, Tom Goldstein, Heng Huang, Mohammad Shoeybi, and Bryan Catanzaro. ODIN: Disentangled reward mitigates hacking in RLHF. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7935–7952. PMLR, 21–27 Jul 2024b. URL https://proceedings.mlr.press/v235/chen24bn.html.

Qiyuan Chen, Hongsen Huang, Jiahe Chen, Qian Shao, Jintai Chen, Hongxia Xu, Renjie Hua, Ren Chuan, and Jian Wu. Learning What Matters: Dynamic Dimension Selection and Aggregation for Interpretable Vision-Language Reward Modeling. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19779–19792, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.906. URL https://aclanthology.org/2026.acl-long.906/.

Jiawei Feng, Jiancan Wu, Xingyu Zhu, Junkang Wu, Xiang Wang, and Xiangnan He. PEA-DPO: Perception-Enhanced Alignment Direct Preference Optimization for MLLMs Alignment. In *Proceedings of the 34th ACM International Conference on Multimedia*, MM '26, New York, NY, USA, November 2026. Association for Computing Machinery. ISBN 979-8-4007-2213-4. doi: 10.1145/3767308.3836113.

Deqing Fu, Tong Xiao, Rui Wang, Wang Zhu, Pengchuan Zhang, Guan Pang, Robin Jia, and Lawrence Chen. TLDR: Token-Level Detective Reward Model for Large Vision Language Models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 45205–45236, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/70217f4d06e57a2395f03b4bc136361a-Paper-Conference.pdf.

Jinlan Fu, Shenzhen Huangfu, Hao Fei, Xiaoyu Shen, Bryan Hooi, Xipeng Qiu, and See-Kiong Ng. CHiP: Cross-modal Hierarchical Direct Preference Optimization for Multimodal LLMs. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 90675–90695, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/e1c73e9595126794186536cfbbed012f-Paper-Conference.pdf.

Leo Gao, John Schulman, and Jacob Hilton. Scaling Laws for Reward Model Overoptimization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022. URL https://mlanthology.org/iclr/2022/hu2022iclr-lora/.

Zeyu Huang, Zihan Qiu, Zili Wang, Edoardo M. Ponti, and Ivan Titov. Post-hoc Reward Calibration: A Case Study on Length Bias. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 37609–37635, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/5d50c76fdf75c24ece568fc84a7125fb-Paper-Conference.pdf.

Zhuoran Jin, Hongbang Yuan, Kejian Zhu, Jiachun Li, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Omni-Reward: Towards Generalist Omni-Modal Reward Modeling with Free-Form Preferences. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 108481–108530, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/b0832cc57899cfd2d3fedeb3f330ba80-Paper-Conference.pdf.

Jiwon Kim and Hyunsoo Yoon. Joint Multimodal Preference Optimization for Fine-Grained Visual-Textual Alignment. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 79–94, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-386-9. doi: 10.18653/v1/2026.findings-eacl.5. URL https://aclanthology.org/2026.findings-eacl.5/.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, Lingpeng Kong, and Qi Liu. VLFeedback: A Large-Scale AI Feedback Dataset for Large Vision-Language Models Alignment. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6227–6246, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.358. URL https://aclanthology.org/2024.emnlp-main.358/.

Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, Lingpeng Kong, and Qi Liu. VL-RewardBench: A Challenging Benchmark for Vision-Language Generative Reward Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24657–24668, June 2025a.

Zichao Li, Xueru Wen, Jie Lou, Yuqiu Ji, Yaojie Lu, Xianpei Han, Debing Zhang, and Le Sun. The Devil Is in the Details: Tackling Unimodal Spurious Correlations for Generalizable Multimodal Reward Models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 36281–36297. PMLR, 13–19 Jul 2025b. URL https://proceedings.mlr.press/v267/li25cw.html.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023. doi: 10.52202/075280-1516. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

Yujie Lu, Dongfu Jiang, Wenhu Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. Wildvision: Evaluating vision-language models in the wild with human preferences. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 48224–48255. Curran Associates, Inc., 2024. doi: 10.52202/079017-1528. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/563991b5c8b45fe75bea42db738223b2-Paper-Datasets_and_Benchmarks_Track.pdf.

OpenAI. Introducing GPT-5.4 mini and nano, March 2026a. URL https://openai.com/index/introducing-gpt-5-4-mini-and-nano/.

OpenAI. GPT-Image-2. OpenAI API model documentation, 2026b. URL https://developers.openai.com/api/docs/models/gpt-image-2.

OpenGVLab. InternVL2-8B. Hugging Face model card, 2024. URL https://huggingface.co/OpenGVLab/InternVL2-8B.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. doi: 10.52202/068431-2011. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Seojeong Park, Jiho Choi, Junyong Kang, Seonho Lee, Jaeyo Shin, and Hyunjung Shim. Mitigating Perceptual Judgment Bias in Multimodal LLM-as-a-Judge via Perceptual Perturbation and Reward Modeling. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Wei Shen, Rui Zheng, Wenyu Zhan, Jun Zhao, Shihan Dou, Tao Gui, Qi Zhang, and Xuanjing Huang. Loose lips sink ships: Mitigating Length Bias in Reinforcement Learning from Human Feedback. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2859–2873, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.188. URL https://aclanthology.org/2023.findings-emnlp.188/.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3008–3021. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning Large Multimodal Models with Factually Augmented RLHF. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.775. URL https://aclanthology.org/2024.findings-acl.775/.

Chenglong Wang, Yang Gan, Yifu Huo, Yongyu Mu, Murun Yang, Qiaozhi He, Tong Xiao, Chunliang Zhang, Tongran Liu, and Jingbo Zhu. RoVRM: A Robust Visual Reward Model Optimized via Auxiliary Textual Preference Data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 25336–25344, 2025a. doi: 10.1609/aaai.v39i24.34721.

Fei Wang, Wenxuan Zhou, James Y. Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mDPO: Conditional Preference Optimization for Multimodal Large Language Models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8078–8088, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.460. URL https://aclanthology.org/2024.emnlp-main.460/.

Xiaokun Wang, Peiyu Wang, Jiangbo Pei, Wei Shen, Yi Peng, Yunzhuo Hao, Weijie Qiu, Ai Jian, Tianyidan Xie, Xuchen Song, Yang Liu, and Yahui Zhou. Skywork-VL Reward: An Effective Reward Model for Multimodal Understanding and Reasoning, 2025b. URL https://arxiv.org/abs/2505.07263.

Shengguang Wu, Fan-Yun Sun, Kaiyue Wen, and Nick Haber. Symmetrical Visual Contrastive Optimization: Aligning Vision-Language Models with Minimal Contrastive Images. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume*

1: Long Papers), pp. 30284–30297, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1462. URL https://aclanthology.org/2025.acl-long.1462/.

Siming Yan, Min Bai, Weifeng Chen, Xiong Zhou, Qixing Huang, and Li Erran Li. ViGoR: Improving Visual Grounding of Large Vision Language Models with Fine-Grained Reward Modeling. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 37–53, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73030-6.

Michihiro Yasunaga, Luke Zettlemoyer, and Marjan Ghazvininejad. Multimodal RewardBench: Holistic Evaluation of Reward Models for Vision Language Models, 2025. URL https://arxiv.org/abs/2502.14191.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. RLHF-V: Towards Trustworthy MLLMs via Behavior Alignment from Fine-grained Correctional Human Feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13807–13816, June 2024a.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, Xiaocheng Feng, Jun Song, Bo Zheng, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. RLAIF-V: Open-Source AI Feedback Leads to Super GPT-4V Trustworthiness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19985–19995, June 2025.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57730–57754. PMLR, 21–27 Jul 2024b. URL https://proceedings.mlr.press/v235/yu24o.html.

Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Ziyu Liu, Shengyuan Ding, Shenxi Wu, Yubo Ma, Haodong Duan, Wenwei Zhang, Kai Chen, Dahua Lin, and Jiaqi Wang. InternLM-XComposer2.5-Reward: A Simple Yet Effective Multi-Modal Reward Model. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 6547–6563, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.340. URL https://aclanthology.org/2025.findings-acl.340/.

YiFan Zhang, HaiHuaYang, Huanyu Zhang, Yang Shi, Zezhou Chen, Haochen Tian, Chaoyou Fu, Kai WU, Bo Cui, Xu Wang, Jianfei Pan, Haotian Wang, Zhang Zhang, and Liang Wang. BaseReward: A Strong Baseline for Multimodal Reward Model. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 113174–113199, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/b7acc4653f48f54bba55f55d2cd10417-Paper-Conference.pdf.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning Modalities in Vision Large Language Models via Preference Fine-tuning, 2024. URL https://arxiv.org/abs/2402.11411.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models, 2025. URL https://arxiv.org/abs/2504.10479.

A Experimental Setup and Reproducibility Details

A.1 Training Data

We construct the training set from three multimodal preference datasets: POVID (Zhou et al., 2024), RLAIF-V (Yu et al., 2025), and RLHF-V (Yu et al., 2024a). POVID contains synthetically constructed preference pairs in which visually grounded responses are contrasted with responses containing injected hallucinations. RLHF-V derives preferences from fine-grained human corrections of hallucinated or visually inconsistent response segments. RLAIF-V provides a substantially larger and more diverse collection of AI-generated preference pairs spanning multiple multimodal tasks and domains.

Before constructing the training pool, we first remove examples overlapping with VL-RewardBench, whose hallucination subset is drawn from POVID, RLAIF-V, and RLHF-V, and then deduplicate the remaining candidates within each source. Multimodal RewardBench and MLLM-as-a-Judge are constructed from separate source datasets and therefore introduce no direct source-level overlap with our three training datasets. We sample the final training mixture from the resulting candidate pools.

All 6,000 training pairs receive standard BT supervision. We apply the additional visual-conditional objectives selectively according to the preference source. Preferences in POVID and RLHF-V are predominantly tied to visual correctness and hallucination-related errors, making them suitable for explicitly supervising the contribution of visual evidence. We therefore partition the 4,180 POVID/RLHF-V training pairs into two disjoint supervision subsets: 3,380 pairs receive VG supervision, while the remaining 800 pairs receive the stronger EF supervision, for which we additionally construct an evidence-flipped image. No training pair receives both VG and EF supervision in the full model. In contrast, RLAIF-V covers broader forms of response quality and can contain valid preference signals attributable to textual quality rather than visual correctness alone; we retain these examples for standard preference learning without applying VG or EF supervision. Details of the EF construction are provided in Appendix B.

A.2 Baselines and Ablations

We compare VCM with standard Bradley–Terry (BT) training and Shortcut-aware (SA) learning (Li et al., 2025b). We additionally consider two ablations, VCM-EF and VCM-VG, to isolate the contributions of evidence-flip and visual-gain supervision. All methods use the same training set and reward-model backbones; their optimization configurations are described in Appendix A.3.

Bradley–Terry. BT is the standard pairwise reward-modeling baseline and is trained only on the original multimodal preference pairs using the objective in Eq. 2.

Shortcut-aware learning. We implement SA following Li et al. (2025b). In addition to the multimodal reward model, SA maintains a separately optimized text-only auxiliary reward model. Let m_θ denote the multimodal preference margin and m_ϕ^T the margin of the text-only auxiliary model, where θ and ϕ denote their respective parameters. Their pairwise losses are

$$\ell_M = -\log \sigma(y m_\theta), \qquad \ell_T = -\log \sigma(y m_\phi^T). \tag{10}$$

SA computes the Shortcut-Failure Coefficient

$$w = \frac{\epsilon/2 - \log \sigma\Big(\mathrm{sg}(y m_\phi^T)\Big)}{\epsilon - \log \sigma\Big(\mathrm{sg}(y m_\phi^T)\Big) - \log \sigma(\mathrm{sg}(y m_\theta))}, \qquad \epsilon = 10^{-7}, \tag{11}$$

and optimizes

$$\mathcal{L}_{\mathrm{SA}} = \mathbb{E}[w\, \ell_M + \ell_T]\,. \tag{12}$$

The margins used to compute w are stop-gradient. The auxiliary text-only model is optimized by ℓ_T, while its prediction dynamically reweights the multimodal BT loss. Only the multimodal reward model is used at inference time.

Table 2: Training-data composition. "Source size" denotes the number of preference pairs in the publicly released dataset. VG and EF denote the disjoint subsets receiving the corresponding visual-conditional supervision.

Dataset	Source size	Train	VG	EF
POVID	17,184	3,200	2,700	500
RLAIF-V	83,132	1,820	–	–
RLHF-V	5,733	980	680	300
Total		6,000	3,380	800

Table 3: Model and training hyperparameters used in the main experiments. Unless specified otherwise, the same configuration is used for all three backbones and training objectives.

Hyperparameter	Setting
LoRA target	All linear projections in LM blocks
LoRA rank / alpha / dropout	32 / 64 / 0.05
Optimizer	AdamW
LoRA learning rate	1×10^{-4}
Reward-head learning rate	1×10^{-3}
Weight decay	0
LR schedule	Cosine, 0.03 warmup ratio
Training length	1 epoch (63 optimizer steps)
Per-device batch size	16 preference pairs
Gradient accumulation	2
Effective global batch size	96 preference pairs
Gradient clipping	4.0
Precision	BF16
Maximum sequence length	2,560
λ_{VG} / λ_{EF}	0.3 / 0.5
γ_{VG} / γ_{EF}	2.0 / 2.0
Text-margin gradient ratio β	0.2
Random seed	42

Ablations.

- **VCM-**EF removes $\mathcal{L}_{\mathrm{EF}}$ and the corresponding edited images while retaining BT and visual-gain supervision. The POVID and RLHF-V examples assigned to EF in the full model are instead included in the VG subset, resulting in 4,180 VG-supervised pairs.
- **VCM-**VG removes $\mathcal{L}_{\mathrm{VG}}$ while retaining BT and evidence-flip supervision.

A.3 Model and Training Details

We train scalar reward models based on Qwen2.5-VL-7B-Instruct (Bai et al., 2025b), Qwen3-VL-8B (Bai et al., 2025a), and InternVL3-8B (Zhu et al., 2025). Each backbone is instantiated as a sequence-classification model with a bias-free linear scalar reward head. The vision encoder and multimodal connector/projector are frozen throughout training. We apply LoRA (Hu et al., 2022) to all linear projections in the language-model transformer blocks and jointly optimize the LoRA parameters and the scalar reward head. The same trainable parameterization is used for BT, VCM, and its ablations; SA additionally maintains an independently optimized text-only auxiliary model with the same LoRA and reward-head configuration.

Training uses three NVIDIA A800 80GB GPUs with distributed data parallelism and gradient checkpointing. For Qwen2.5-VL and Qwen3-VL, images are processed with a maximum of 512 visual tokens and a pixel budget between 3,136 and 401,408 pixels; InternVL3 uses at most two dynamic image patches. We use the final checkpoint after the single training epoch for evaluation.

Table 4: Evaluation benchmarks and subsets used in the main experiments.

Benchmark	Subset	Metric	Size
Held-out	POVID/RLAIF-V/RLHF-V	Pair acc.	$3{,}000 \times 3$
VLF	Non-tied response pairs	Instr. acc.	$1{,}000 \times 3$ instr.
WildVision	Single-turn, strict preference	Pair acc.	$1{,}000 \times 3$
VLRB	Hallucination	Pair acc.	749
MMRB	VQA + Correctness	Pair acc.	1,823
MLLM-Pair	Pairwise benchmark	Pair acc.	6,165

A.4 Evaluation Setup

We evaluate the reward models on both held-out preference data from the training sources and external multimodal reward-model benchmarks. Except for VLFeedback (VLF), the primary metric is pairwise accuracy, where a prediction is correct if the preferred response receives a strictly higher reward than the rejected response. For benchmarks with three evaluation rolls, we report the mean and sample standard deviation across rolls; fixed public benchmarks are evaluated on the full selected subset. Table 4 summarizes the evaluation data used in the main experiments.

Held-out evaluation. We construct three fixed held-out rolls, each containing 3,000 preference pairs from POVID, RLAIF-V, and RLHF-V and disjoint from the training pool. Accuracy is computed over all pairs in each roll and then averaged across the three rolls.

VLFeedback. We construct three disjoint evaluation rolls with 1,000 instructions each. We remove malformed examples and overlaps with the training, held-out, and other external evaluation sets before sampling. Each instruction contains four candidate responses scored along Helpfulness, Ethical Considerations, and Visual Faithfulness. We average the three provided annotation scores for each response, enumerate the six unordered response pairs, and discard pairs with tied preference scores. This yields 4,953, 4,914, and 4,967 non-tied pairs in the three rolls, respectively. Because the number of valid pairs varies across instructions, we first compute pairwise accuracy within each instruction and then macro-average over the 1,000 instructions; this instruction-level accuracy is the VLF metric reported in the main results.

WildVision-Battle. We retain single-turn battles with a strict winner, remove duplicate conversations and duplicate image–question instances, and exclude overlap with VL-RewardBench before constructing three fixed rolls of 1,000 pairs. Pairwise accuracy is computed independently for each roll and then averaged across the three rolls.

Fixed public benchmarks. For VL-RewardBench (VLRB), we evaluate on its hallucination subset, which overlaps in source datasets with our training data; overlapping examples are removed before the training set is sampled, leaving 749 evaluation pairs. For Multimodal RewardBench (MMRB), we combine the VQA and general-correctness subsets, containing 1,200 and 623 pairs respectively, and report micro-averaged accuracy over all 1,823 pairs. For MLLM-as-a-Judge Pair (MLLM-Pair), we evaluate on the complete pairwise set of 6,165 examples. MMRB and MLLM-Pair are constructed from data sources separate from POVID, RLAIF-V, and RLHF-V and therefore have no direct source-level overlap with the training mixture.

A.5 Best-of-N Evaluation

Following Li et al. (2025b), we evaluate whether the learned reward models improve test-time response selection on three free-form multimodal benchmarks: MM-Vet (Yu et al., 2024b), LLaVA-Bench-in-the-Wild (Liu et al., 2023), and MMHal-V (Sun et al., 2024; Amirloo et al., 2024). MM-Vet evaluates integrated multimodal capabilities such as recognition, OCR, knowledge, and reasoning; LLaVA-Bench-in-the-Wild contains open-ended visual instructions requiring helpful and grounded responses; MMHal-V is the image-conditioned variant of MMHal-Bench and focuses on hallucination-sensitive visual faithfulness. We evaluate 218 MM-Vet examples, 60 LLaVA-Bench-in-the-Wild examples, and 96 MMHal-V examples, for a total of 374 image–query pairs.

For each image–query pair, a frozen InternVL2-8B policy generates a shared pool of 64 candidate responses. Candidate generation uses temperature 1.0, top-p 0.95, a maximum of 1,024 new tokens, and repetition penalty 1.0. The resulting 23,936 candidates are fixed before reward-model evaluation and shared across BT, SA, and VCM. Each reward model independently scores all 64 candidates using the original image and query, and BoN selection chooses the candidate with the highest reward among the considered subset.

We use a fixed GPT-5.4 mini multimodal evaluator with temperature 0 to assign an external quality score to every candidate response before BoN aggregation. For MM-Vet, the evaluator assigns a correctness score in $[0, 1]$. For LLaVA-Bench-in-the-Wild, it assigns a $[1, 10]$ score based on helpfulness, correctness, relevance, and grounding in the visual input. For MMHal-V, it assigns a $[0, 6]$ score emphasizing visual faithfulness and the absence of hallucinated content. These evaluator scores are computed once and shared across all reward models, so the compared methods differ only in how they rank and select responses.

We report $N \in \{1, 2, 4, 8, 16, 32, 64\}$. Rather than approximating BoN performance using repeatedly sampled subsets, we compute its exact expectation over uniformly sampled subsets of size N from the fixed pool of 64 candidates. Let $a_{(k)}$ denote the candidate with the k-th highest reward and let $s(a_{(k)})$ denote its external evaluator score. The probability that $a_{(k)}$ is the highest-reward candidate contained in a uniformly sampled subset of size N is

$$p_{k,N} = \frac{\binom{64-k}{N-1}}{\binom{64}{N}}. \tag{13}$$

The expected BoN score for an image–query pair is therefore

$$S_N = \sum_{k=1}^{64} p_{k,N} s(a_{(k)}). \tag{14}$$

We average S_N over examples within each benchmark to obtain the reported MM-Vet, LLaVA-Bench-in-the-Wild, and MMHal-V curves. For visualization in Figure 6, MM-Vet and LLaVA-Bench-in-the-Wild scores are reported as percentages of their maximum evaluator scores, i.e., the native $[0, 1]$ and $[1, 10]$ scores are multiplied by 100 and 10, respectively. MMHal-V retains its native $[0, 6]$ scale.

B Evidence-Flip Sample Construction and Quality Inspection

Evidence-flip (EF) samples are constructed from existing multimodal preference pairs by modifying only the visual evidence while keeping the question and both candidate responses unchanged. Given an original pair whose preference is supported by the image, the goal is to make the smallest local modification to the decisive visual evidence such that the alternative response becomes preferable under the edited image. We initially construct 1,000 original–edited image pairs, consisting of 500 POVID, 200 RLAIF-V, and 300 RLHF-V examples. The final VCM models use the 800 EF pairs from POVID and RLHF-V, consistent with the visually grounded preference sources used for visual-conditional supervision; the RLAIF-V edits are not used in the reported EF objective.

Construction. Each EF example is associated with one original training pair containing an image, a question, a preferred response, and a rejected response. Candidate pairs are selected from the training pool and checked to contain a valid image and two distinct responses. Each selected pair is associated with exactly one edited image. The question and candidate responses are copied without modification, and only the image is replaced in the EF state. Thus, an original and its EF counterpart share the same textual context while differing in the visual evidence that determines the intended preference.

Image editing. We generate edited images using `gpt-image-2` (OpenAI, 2026b). The model receives the original image together with the question, the current preferred response, and the alternative response. We use the same pair-conditioned instruction for all examples:

```
Make the smallest possible local edit to the image so that, for
the user question below, the alternative response becomes a better
answer than the current gold response.

Preserve the image outside the minimum region needed for this
preference reversal.

User question:
{question}

Current gold response:
{gold_response}

Alternative response:
{alternative_response}
```

Each request generates a single edited image with medium generation quality at 1024×1024 resolution. The generated image is resized to the dimensions of the corresponding original image before being used for training. We do not manually modify or composite the generated images. Requests that fail to return a valid image are discarded and replaced with another unused candidate from the same source. The edit type is not predefined; instead, it is determined by the visual disagreement between the two responses and can involve object presence or identity, object count, attributes, spatial relations, or visible text.

Manual quality inspection. We manually inspect 100 randomly sampled EF pairs from the final training set. We evaluate two properties directly related to the construction objective. An edit is counted as a *local-edit success* when the intended visual evidence is changed while the relevant non-target content remains substantially preserved; 97% of the inspected samples satisfy this criterion. We further evaluate whether the edited image actually reverses the preference between the unchanged candidate responses, obtaining an *evidence-flip success rate* of 88%. We additionally check for severe generation artifacts during inspection and do not observe systematic failures that would materially affect the preference judgment.

Example. Figure 7 shows a representative EF pair from RLHF-V. The original image depicts a conventional road intersection, while the edited image introduces a roundabout in the foreground. The question and the two responses remain unchanged; the local visual intervention changes the decisive evidence from a **regular intersection** to a **roundabout intersection**, reversing the intended preference.

Controlling for edited-image exposure. A remaining concern is whether improvements could arise simply because the model is exposed to additional edited images during training. Appendix E.1 provides a matched-exposure comparison in which BT, SA, and VCM all receive the same 800 edited and preference-reversed EF samples through ordinary pairwise preference optimization. The resulting comparison shows that additional EF-image exposure alone does not consistently reproduce the behavior of VCM, providing complementary evidence that the benefit is associated with explicitly supervising how preference margins respond to controlled visual changes rather than merely adding counterfactual images to the training set.

C Additional Preliminary Analyses

This section provides additional details and results for the preliminary study in Section 3. We use the same experimental setting to examine two questions: whether pretrained VLM representations already contain text-only preference signals, and how these signals evolve under standard Bradley–Terry training.

C.1 Preliminary Study Setup

We conduct the preliminary study on POVID, RLAIF-V, and RLHF-V using Qwen2.5-VL-7B-Instruct and InternVL3-8B. For each source, we sample 3,000 preference pairs for training and 1,000 disjoint pairs for evaluation, resulting in a balanced 9,000-pair training set and a 3,000-pair



Original image



Evidence-flipped image

Question: What kind of intersection are the cars on?

Response A: The cars are at a **regular intersection**. A regular intersection is the most common type of intersection, formed by two roads crossing on the same level. Regular intersections usually have traffic signals or traffic signs to guide the movement of vehicles and pedestrians. The presence of traffic signals effectively controls vehicle flow, ensuring traffic safety and smoothness.

Response B: The cars are moving at a **roundabout intersection**. A roundabout intersection is a special type of intersection, with roads in a circular shape and usually multiple entrances and exits. Vehicles must yield to other vehicles before entering, and traffic flows in a clockwise direction around a central island. Roundabout intersections are widely used in Europe and other parts of the world to manage traffic flow and reduce the risk of serious accidents.

Preference: Original: **A** $\succ$ **B** Evidence-flipped: **B** $\succ$ **A**.

Figure 7: Representative evidence-flip example from RLHF-V. The question and candidate responses remain fixed. A local edit changes the decisive road structure from a **regular intersection** to a **roundabout intersection**, reversing the intended preference.

Table 5: Data splits used in the preliminary study.

Split	POVID	RLAIF-V	RLHF-V	Total
Training	3,000	3,000	3,000	9,000
Test	1,000	1,000	1,000	3,000
BT monitor	256	256	256	768

test set. To monitor the dynamics of BT training, we additionally fix 256 test pairs from each source, yielding a 768-pair monitor set that is never used for optimization.

Frozen representation probes. Following the setup in Section 3.1, we extract the last non-padding response-token representation under multimodal and text-only inputs. In addition to the final representation reported in the main text, we evaluate intermediate readouts at language-model depths 12, 16, 20, and 24. For each readout, we fit an intercept-free L2-regularized logistic BT probe with $C = 1$ using LBFGS for up to 2,000 iterations. Representation standardization is estimated from the training split only. We also include a random-label control by applying a fixed random sign to the training pair features while retaining the original test labels.

BT training dynamics. Following Section 3.2, we train the reward model for one epoch using only the standard multimodal BT objective. With a global batch size of 96 pairs, training consists of 94 optimizer steps. We evaluate checkpoints at steps 0, 6, 18, 30, 45, 60, 76, and 94 on the fixed 768-pair monitor, recording multimodal and text-only pairwise accuracy, signed preference margins, and their Spearman correlation. Text-only inference is used only for analysis and does not contribute gradients during BT training.

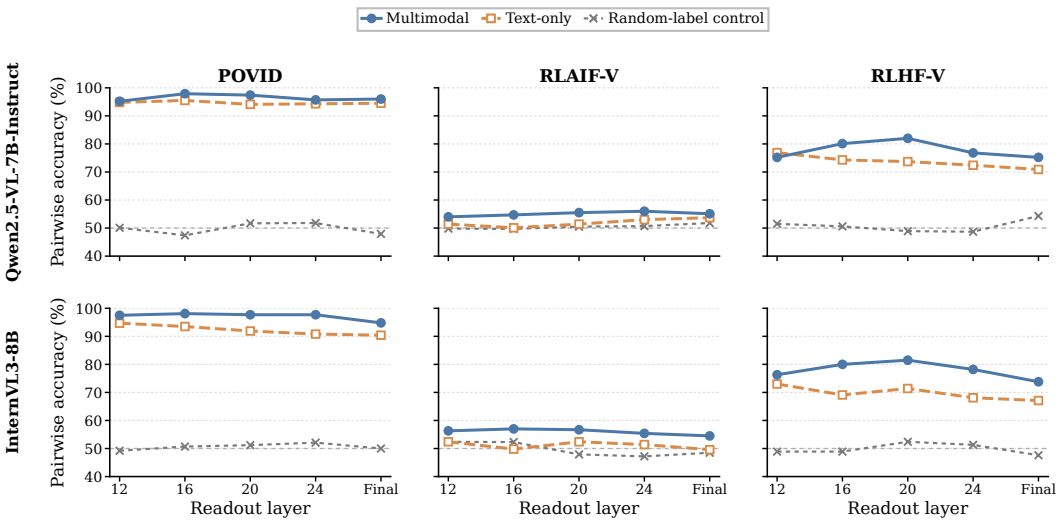


Figure 8: Layer-wise preference-probe accuracy in frozen pretrained representations. Multimodal and text-only probes are evaluated at intermediate language-model readouts and the final normalized representation for Qwen2.5-VL-7B-Instruct and InternVL3-8B. Random-label controls use the same probe-fitting procedure with shuffled training labels. Text-only preference predictability remains stable across readout depth on POVID and RLHF-V, while RLAIF-V and the random-label controls remain close to chance.

C.2 Text-Only Preference Signals in Frozen Representations

Figure 8 extends the final-readout results in Section 3.1 to intermediate language-model representations. Across both backbones, the source-level pattern is stable over readout depth. For Qwen2.5-VL-7B, text-only probe accuracy ranges from 94.1–95.5% on POVID, 50.1–53.7% on RLAIF-V, and 70.9–76.9% on RLHF-V. For InternVL3-8B, the corresponding ranges are 90.4–94.7%, 49.6–52.4%, and 67.1–73.0%, respectively. Multimodal probes generally remain stronger than their text-only counterparts, while preserving the same ordering across data sources.

The random-label controls remain close to chance across readout positions, indicating that the above-chance text-only predictability is not an artifact of the linear-probe fitting procedure. More importantly, the signal is not confined to the final representation: POVID and RLHF-V exhibit substantial text-only preference predictability across multiple intermediate layers, whereas RLAIF-V remains close to chance. This suggests that the source-dependent textual preference structure observed in the main text is already broadly represented within the pretrained VLM rather than emerging only at the final readout.

C.3 Preference Dynamics under Standard BT Training

Figure 9 complements the accuracy trajectories in Section 3.2 by tracking signed preference margins and their alignment during standard BT training. For both backbones, the multimodal signed margin increases as optimization progresses, while the text-only signed margin remains positive and exhibits an overall upward shift, particularly for InternVL3-8B. This shows that the improvement in pairwise prediction is accompanied by a strengthening of preference structure that remains accessible without the image.

At the same time, the correlation between multimodal and text-only margins increases throughout training, and the learned text-only margin becomes increasingly aligned with the text-only preference direction identified by the frozen probe in Section C.2. These results do not imply that BT fails to learn visual information; rather, they show that standard pairwise optimization can reinforce pre-existing textual preference structure alongside improvements in multimodal preference prediction.

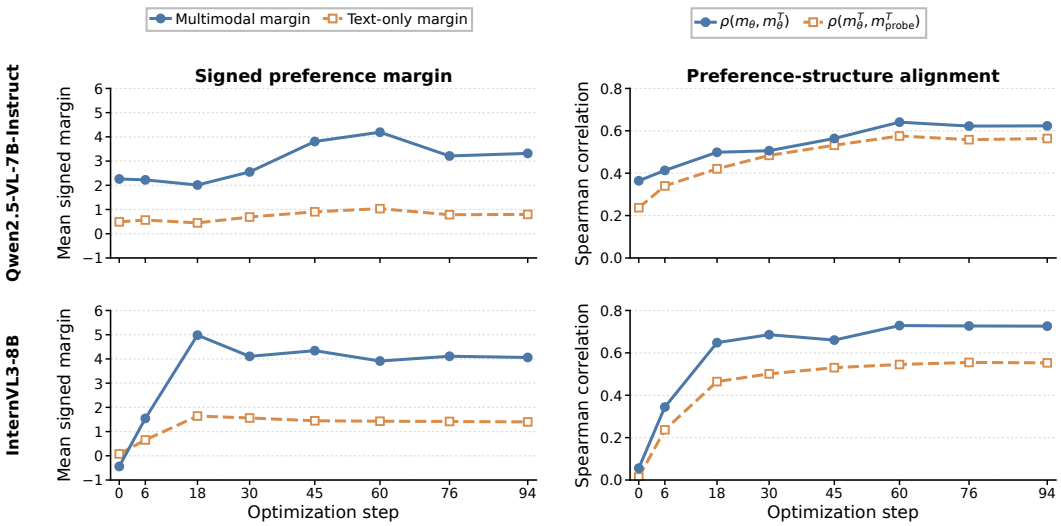


Figure 9: Preference-margin dynamics under standard BT training. Left: mean signed multimodal and text-only preference margins over optimization. Right: Spearman correlation between multimodal and text-only margins, and between the learned text-only margin and the frozen text-only probe margin. Standard BT increases the overall preference margin while simultaneously strengthening alignment with pre-existing text-only preference structure.

C.4 Image-Mismatch Diagnostic

We further examine whether the preference structure learned by standard BT depends on the correspondence between the image and the textual context. At the final BT checkpoint, we construct a mismatch condition on the fixed probe test set by replacing each image with another image from the same source and data bucket, while keeping the question and both candidate responses unchanged. The replacement is performed as a one-to-one derangement without fixed points or identical images. We denote the resulting preference margin by m_θ^{mis} and compare it with the original multimodal margin and the text-only margin m_θ^T.

As shown in Figure 10, replacing the matched image substantially reduces pairwise accuracy, bringing performance closer to the text-only setting for both backbones. Meanwhile, the mismatch margin m_θ^{mis} remains positively correlated with the text-only margin m_θ^T, with an overall Spearman correlation of approximately 0.64–0.66. Although the mismatched image still changes the reward values, the resulting preference behavior becomes substantially closer to the text-only regime once the original visual evidence is no longer informative. This pattern is consistent with the model falling back on preference structure available from text alone when the matched visual evidence is disrupted.

D An Optimization Perspective on Bradley–Terry Training

Section 4.1 introduces the exact reference-state decomposition $m_\theta = m_\theta^T + m_\theta^\times$, where m_θ^T is the preference margin obtained from the same reward model without the image and $m_\theta^\times = m_\theta - m_\theta^T$ is the incremental margin induced by multimodal conditioning. This decomposition is an algebraic identity rather than an assumption that the reward model internally contains independently parameterized textual and visual pathways. It nevertheless provides a convenient functional view for examining what standard Bradley–Terry supervision constrains.

BT constrains only the aggregate margin. For a single preference example, the BT loss can be written in the margin coordinates $(m_\theta^T, m_\theta^\times)$ as

$$\ell_{\text{BT}}(m_\theta^T, m_\theta^\times) = \log\big(1 + \exp\big[-y(m_\theta^T + m_\theta^\times)\big]\big)\,. \tag{15}$$

Therefore, for any scalar δ,

$$\ell_{\text{BT}}\left(m_\theta^T + \delta,\, m_\theta^\times - \delta\right) = \ell_{\text{BT}}\left(m_\theta^T,\, m_\theta^\times\right). \tag{16}$$

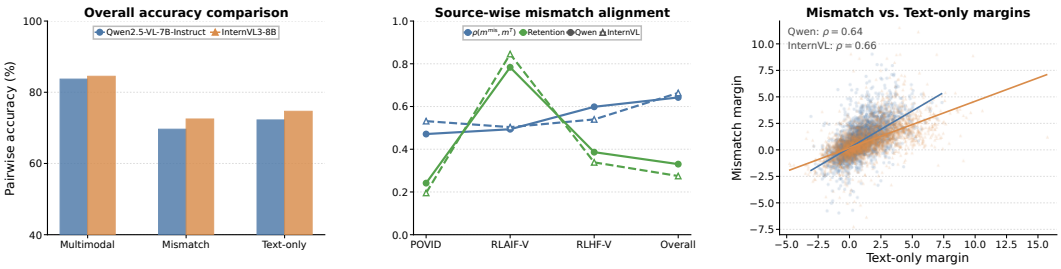


Figure 10: Image-mismatch diagnostic for BT-trained reward models. Left: pairwise accuracy under the original multimodal input, mismatched images, and text-only input. Middle: source-wise Spearman correlation between mismatch and text-only margins together with mismatch-to-multimodal margin retention. Right: per-example mismatch margins against text-only margins on the complete probe test set. Replacing the matched image degrades preference accuracy while the resulting margins remain partially aligned with the text-only preference structure.

In these functional margin coordinates, BT is insensitive to how a fixed total preference margin is allocated between the text-only reference margin and the multimodal increment. Equivalently, the two partial derivatives are identical:

$$\frac{\partial \ell_{\mathrm{BT}}}{\partial m_\theta^T} = \frac{\partial \ell_{\mathrm{BT}}}{\partial m_\theta^\times} = -y\,\sigma\big[-y(m_\theta^T + m_\theta^\times)\big]\,. \tag{17}$$

The corresponding Hessian takes the form

$$\nabla^2 \ell_{\mathrm{BT}} = c\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \qquad c = \sigma(y m_\theta)\sigma(-y m_\theta) \geq 0, \tag{18}$$

and therefore has a zero-curvature direction $(1, -1)$ that changes the component allocation while preserving the total margin. This observation concerns the BT objective in functional margin space; because m_θ^T and $m_\theta^\times$ are coupled functions of the same model parameters, it does not imply an exact flat direction in neural-network parameter space.

A shortcut-compatible limiting case. This view also illustrates why text-only shortcuts can support arbitrarily low BT loss when preference labels are predictable from text. Suppose that, for a set of training examples, the model can realize a text-only margin satisfying

$$y_i m_i^T \geq \epsilon > 0. \tag{19}$$

If the multimodal increment vanishes, $m_i^\times = 0$, and the text-only margin can be scaled by $\alpha > 0$, then

$$\ell_{\mathrm{BT},i} = \log\big(1 + \exp(-\alpha y_i m_i^T)\big) \leq \log(1 + \exp(-\alpha\epsilon)) \xrightarrow[\alpha\to\infty]{} 0. \tag{20}$$

Thus, under this idealized separability condition, BT can drive the preference loss arbitrarily close to zero without requiring an additional margin from visual conditioning. Real multimodal preference data need not satisfy this condition globally; however, whenever a subset of examples is already well predicted from text, the standard BT objective provides no separate requirement that their correct margins also be supported by matched visual evidence. This simple case is consistent with the empirical dynamics in Section 3 and Appendix C, where text-only preference signals are present before reward-model training and become increasingly aligned with the learned multimodal preference structure under BT.

Relation to shortcut-aware reweighting. Shortcut-aware learning changes the importance of examples according to a separately trained text-only auxiliary model. For the multimodal branch, its update remains a detached sample-wise reweighting of the BT loss,

$$\mathcal{L}_{\mathrm{SA}} = \mathbb{E}_i\left[w_i\,\ell_{\mathrm{BT},i}\right] + \mathcal{L}_{\mathrm{aux}}, \tag{21}$$

where w_i is determined from the text-only auxiliary model. At a fixed weight w_i, the multimodal objective still depends on the aggregate margin $m_{\theta,i} = m_{\theta,i}^T + m_{\theta,i}^\times$ rather than directly constraining $m_{\theta,i}^\times$. Reweighting can therefore reduce the optimization pressure contributed by examples that are easily explained from text, but it does not explicitly specify how the preference margin of an individual example should change under multimodal conditioning.

Table 6: Results under matched evidence-flip data exposure. BT, SA, and VCM are all trained with the same 800 edited and preference-reversed EF samples included in pairwise preference optimization.

Backbone	Method	Held-out↑	VLF↑	VLRB-H↑	MMRB↑	MLLM-Pair↑
Qwen2.5-VL-7B	BT	79.52±0.44	79.02±0.81	82.24	75.15	59.17
	SA	79.70±0.34	72.15±0.58	76.37	75.43	59.68
	VCM	81.18±0.68	78.80±1.12	84.11	76.91	64.79
Qwen3-VL-8B	BT	81.71±1.06	76.28±0.91	86.92	76.96	62.01
	SA	80.63±1.22	71.99±0.91	78.64	78.72	61.80
	VCM	81.39±0.95	80.52±0.64	86.78	79.05	63.00
InternVL3-8B	BT	81.76±0.80	75.89±0.86	85.31	74.60	60.24
	SA	80.08±0.42	73.88±0.48	74.77	74.66	59.35
	VCM	80.46±0.49	77.81±0.84	78.10	67.09	60.32

Visual-conditional supervision. VCM introduces the missing component-specific constraints directly. Visual-gain supervision operates on

$$m_\theta - m_\theta^T = m_\theta^\times, \tag{22}$$

while evidence-flip supervision compares two visual states with identical questions and responses, such that the shared text-only reference margin cancels from the cross-state difference. From this perspective, VG and EF do not replace the standard BT objective; instead, they supplement its supervision by constraining aspects of the preference margin that are left unspecified by pairwise preference labels alone.

E Additional Evaluation and Generalization Results

E.1 Controlling for Evidence-Flip Data Exposure

To examine whether the gains of visual-conditional supervision can be attributed simply to additional exposure to edited images, we construct an exposure-controlled setting in which BT, SA, and VCM are all trained with the same 800 evidence-flipped samples as additional preference pairs. Each edited image is paired with the reversed preference label and participates in the corresponding BT objective. VCM additionally retains its visual-gain and evidence-flip supervision. This comparison isolates whether treating EF samples as additional preference data alone can account for the improvements obtained from explicitly supervising visual-conditional margins.

The exposure-controlled comparison shows that simply incorporating evidence-flipped samples into ordinary preference training does not consistently reproduce the behavior of VCM. The clearest differences appear on VLF and MLLM-Pair for the Qwen backbones, while the InternVL results are more mixed. Overall, the results suggest that the effect of VCM is not explained solely by additional exposure to edited counterfactual images.

E.2 Generalization to Smaller Backbones

We further evaluate VCM on Qwen3-VL-4B-Instruct and InternVL3-2B using the same training data, optimization schedule, and VG/EF supervision protocol as the 7B/8B models in the main experiments.

On Qwen3-VL-4B, VCM improves VLFeedback, MMRB, and MLLM-Pair, while remaining comparable on Held-out and lower on VLRB-H. On InternVL3-2B, the improvements are smaller but remain positive on VLFeedback, MMRB, and MLLM-Pair, with BT performing better on Held-out and VLRB-H. Overall, the smaller-backbone results show that the effect of visual-conditional supervision is not limited to the 7B/8B setting.

Table 7: Results on smaller reward-model backbones. Held-out and VLFeedback report the mean across three fixed evaluation rolls.

Backbone	Method	Held-out↑	VLFeedback↑	VLRB-H↑	MMRB↑	MLLM-Pair↑
Qwen3-VL-4B	BT	82.39±0.62	74.67±0.34	91.05	72.68	58.46
	VCM	82.53±0.20	79.32±0.69	88.52	76.69	61.98
InternVL3-2B	BT	78.23±0.25	75.46±0.33	86.52	60.50	57.27
	VCM	77.60±0.58	75.90±0.48	85.05	64.51	58.07

Table 8: Results on additional VL-RewardBench and Multimodal RewardBench categories.

		VL-RewardBench		Multimodal RewardBench			
Backbone	Method	General↑ $N = 181$	Reasoning↑ $N = 317$	Reasoning↑ $N = 1{,}096$	Knowledge↑ $N = 630$	Preference↑ $N = 654$	Safety↑ $N = 508$
Qwen2.5-VL-7B	BT	51.38	54.89	65.97	57.94	55.96	81.10
	SA	46.96	59.62	68.89	60.63	53.82	87.20
	VCM	46.96	58.04	56.84	62.86	60.40	90.35
Qwen3-VL-8B	BT	51.38	62.78	69.25	66.67	50.31	38.78
	SA	50.83	66.25	70.53	68.89	49.85	50.20
	VCM	48.62	70.66	67.88	67.78	57.65	51.97
InternVL3-8B	BT	54.70	59.62	60.95	59.52	47.86	71.06
	SA	53.04	55.21	66.51	58.25	47.09	76.57
	VCM	55.80	57.10	66.51	59.37	58.26	73.23

E.3 Results across Additional Benchmark Categories

We further evaluate the main 7B/8B reward models on benchmark categories not reported in the main table. For VL-RewardBench, we include the General and Reasoning subsets. For Multimodal RewardBench, we report the Reasoning, Knowledge, Preference, and Safety subsets.

The category-level results are mixed across benchmarks and backbones. VCM shows clear improvements on the MMRB preference subset across all three backbones, and also improves several knowledge, safety, and VLRB reasoning settings. Other categories favor BT or SA, indicating that the effect of visual-conditional supervision depends on the type of preference signal being evaluated.